BIOSTATISTICĂ pentru EPIDEMIOLOGIE și SĂNĂTATE PUBLICĂ FOLOSIND R

Bertram KC Chan


SPRINGER PUBLISHING COMP

Biostatistica pentru Epidemiologie și Public Sănătate folosind R

Bertram KC Chan, PhD, PE, și-a terminat studiile secundare în Sydney, Australia, după ce a promovat certificatul de absolvire a statului New South Wales (matriculare universitară) cu rezultate excelente la matematică și la fizică și chimie. Apoi a absolvit atât o licență în științe în inginerie chimică, cu distincții de primă clasă (summa cum laude), cât și un master în științe inginerești în inginerie nucleară la Universitatea din New South Wales și un doctorat în inginerie la Universitatea din Sydney. . Aceasta a fost urmată de 2 ani de activitate de cercetare la Institutul de Cercetare al Comisiei pentru Energie Atomică din Australia și 2 ani de bursă postdoctorală a Comisiei pentru Energie Atomică din Canada la Universitatea din Waterloo, Canada.

A urmat studii postuniversitare suplimentare la Universitatea din New South Wales, la Universitatea Americană din Beirut și la Universitatea Stanford, în statistică , informatică și matematică pură și aplicată (algebră abstractă, teoria automatelor, analiză numerică etc.) , și în electronică și inginerie electromagnetică.

Cariera sa profesională include peste 10 ani de experiență de predare și cercetare la nivel universitar în mai multe instituții, inclusiv cercetător asociat în analiză biomedicală și statistică, Secțiunea de biologie perinatală, Departamentul de ginecologie, Facultatea de Medicină a Universității din California de Sud, Universitatea Orientului Mijlociu și San Jose. State University și posturi de cercetare industrială la Lockheed Missile & Space, Apple, Hewlett-Packard și o companie start-up (Foundry Networks) în fabricarea de hardware și software pentru Internet: comutatoare și routere gigahertz . În ultimii ani, el a susținut activitatea de biostatistică a programului de cercetare Adventist Health Studies II la Școala de Medicină a Universității Loma Linda (LLU) din California și a consultat ca lector pe forum timp de câțiva ani la Școala de Sănătate Publică LLU (biostatistică). , epidemiologie și medicina populației). Prelegerile LLU au făcut parte din acest manual. În aceste prelegeri, dr. Chan a introdus utilizarea limbajului de programare R și a conceput aceste prelegeri pentru elementele biostatistice pentru cursurile din programele MPH, MsPH, DrPH și doctorat, cu referire specială la epidemiologie în special și sănătatea publică și populație. medicina lațiune în general.

Dr. Chan deține trei brevete din SUA și a publicat peste 30 de lucrări de cercetare și a fost autor a 10 cărți de matematică educațională. Este un inginer profesionist înregistrat (PE) în statul California.

Biostatistica pentru Epidemiologie și Public Sănătate folosind R

Bertram KC Chan, PhD, PE

Springer y TPUBLISHING COMPANY
NEW YORK

Copyright © 2016 Springer Publishing Company, LLC

Toate drepturile rezervate.

Nicio parte a acestei publicaţii nu poate fi reprodusă, stocată într-un sistem de recuperare sau transmisă sub nicio formă sau prin orice mijloc, electronic, mecanic, fotocopiere, înregistrare sau altfel, fără permisiunea prealabilă a Springer Pub lishing Company, LLC sau autorizaţia. prin plata taxelor corespunzătoare către Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400, fax 978-646-8600, info@copyright.com sau pe Web la www.copyright.com.

Springer Publishing Company, LLC
11 West 42nd Street
New York, NY 10036
www.springerpub.com

Editor Achiziţii : Nancy S. Hale
Compozitie : Exeter Premedia Services Private Ltd.

ISBN: 978-0-8261-1025-1
e-book ISBN: 978-0-8261-1026-8

Manualul instructorului ISBN: **978-0-8261-3279-6. Instructorii calificaţi pot solicita suplimente prin e-mail textbook@springerpub.com**

Materiale pentru studenţi: **Ghid de studiu pentru studenţi (ISBN: 978-0-8261-3278-9) şi *Capitolul suplimentar:* Aplicaţii la nivel de cercetare ale lui R (ISBN: 978-0-8261-3249-9) sunt disponibile la springerpub .com/chan-biostatistics**

15 16 17 18 / 5 4 3 2 1

Autorul şi editorul acestei Lucrări au depus toate eforturile pentru a utiliza surse considerate a fi de încredere pentru a oferi informaţii care sunt corecte şi compatibile cu standardele general acceptate la momentul publicării. Autorul şi editorul nu vor fi răspunzători pentru nicio daune speciale, consecutive sau exemplare care rezultă, în întregime sau parţial, din utilizarea de către cititori a informaţiilor conţinute în această carte sau încrederea în acestea. Editorul nu are nicio responsabilitate pentru persistenţa sau acurateţea URL-urilor pentru site-urile web externe sau ale terţilor la care se face referire în această publicaţie şi nu garantează că orice conţinut de pe astfel de site-uri web este sau va rămâne corect sau adecvat.

Datele de catalogare în publicaţie ale Bibliotecii Congresului

Chan, BKC (Bertram Kim-Cheong), autor.
Biostatistică pentru epidemiologie şi sănătate publică folosind R / Bertram KC Chan.
p. ; cm.
Include referinţe bibliografice şi index.
ISBN 978-0-8261-1025-1 — ISBN 978-0-8261-1026-8 (carte electronică)
I. Titlu.
[DNLM: 1. Biostatistică—metode. 2. Epidemiologie. 3. Limbaje de programare.
4. Sănătate Publică. WA 950]
R853.S7
610.1'5195—dc23

2015024102

Reduceri speciale la cantităţile vrac ale cărţilor noastre sunt disponibile pentru corporaţii, asociaţii profesionale, companii farmaceutice, organizaţii de îngrijire a sănătăţii şi alte grupuri eligibile. Dacă sunteţi interesat de o carte personalizată, inclusiv capitole din mai mult de unul dintre titlurile noastre, vă putem oferi şi acel serviciu.

Pentru detalii, vă rugăm să contactaţi:
Departamentul special de vânzări, Springer Publishing Company, LLC
11 West 42nd Street, 15th Floor, New York, NY 10036-8002
Telefon: 877-687-7476 sau 212-431-4370; Fax: 212-941-7842

E-mail: sales@springerpub.com

Tipărit în Statele Unite ale Americii de Bradford & Bigelow.

Cuprins

Prefață xi
Distribuiți Biostatistică pentru Epidemiologie și Sănătate Publică Folosind R

1. INTRODUCERE 1
1.1 *Medicină, Medicină Preventivă, Sănătate Publică și Epidemiologie 1* Medicina *1*
Medicina preventiva si sanatate publica *1*
Sănătate publică și epidemiologie *2*
Întrebări de revizuire pentru Secțiunea 1.1 *2*
1.2 *Sănătatea personală și sănătatea publică 3*
Sănătatea personală versus sănătatea publică *4*
Întrebări de revizuire pentru Secțiunea 1.2 *5*
1.3 *Cercetare și măsurători în EPDM și PH 6*
EPDM: Știința de bază a PH *7*
Principalele funcții epidemiologice *9*
Cauza bolilor *14*
Măsurarea expunerii în epidemiologie *19*
Probleme suplimentare *20*
Întrebări de revizuire pentru Secțiunea 1.3 *20*
1.4 *BIOS și EPDM 21*
Întrebări de revizuire pentru Secțiunea 1.4 *23*
Referințe 23

2. CERCETARE ȘI PROIECTARE ÎN EPIDEMIOLOGIE ȘI SĂNĂTATE PUBLICĂ 25
Introducere 25
2.1 *Cauzalitate și asociere în epidemiologie și sănătate publică 26*
Criteriile Bradford-Hill pentru cauzalitate și asociere în epidemiologie *26*
Interpretare juridică folosind epidemiologie *28*
Apariția bolii *29*
Întrebări de revizuire pentru Secțiunea 2.1 *32*
2.2 *Cauzalitate și inferență în epidemiologie și sănătate publică 33*
Diagramele lui Rothman pentru cauzarea suficientă a bolilor *33*
Inferențe cauzale *37*
Utilizarea criteriilor cauzale *38*
Judecarea dovezilor științifice *38*
Întrebări de revizuire pentru Secțiunea 2.2 *39*
2.3 *Bazele biostatistice ale inferenței 40*
Moduri de inferență *40*
Niveluri de măsurare *40*
BIOS frecventist în EPDM *41*
Intervalele de încredere în epidemiologie și sănătate publică *42*
Interval Bayesian credibil *48*
Întrebări de revizuire pentru Secțiunea 2.3 *49*

2.4 *BIOS în EPDM şi PH* 50
Aplicaţii ale BIOS 50
BIOS în EPDM şi PH 51

Prelucrarea și analiza datelor epidemiologice de bază *52*
Analiza datelor epidemiologice *56*
Folosind R *67*
Evaluarea unei singure măsuri de apariție *67*
Numărul de Poisson (incidența) și datele privind rata *68*
Date binomiale de risc și prevalență *68*
Evaluarea a două măsuri de apariție — Comparația riscului: raportul de risc și riscul atribuibil *69*
Compararea a două estimări de rate: Rate Rate *rr 70*
Compararea a două estimări de risc: raportul de risc *RR și* raportul cotelor de boală (morbiditate) *DOR 70*
Compararea a două estimări de șanse de la caz-control: Studiul epidemiologic privind vaccinul antipoliomielită Salk *72*
Întrebări de revizuire pentru Secțiunea 2.4 *72*
Exerciții pentru capitolul 2 73
Folosirea teoriei probabilităților *73*
Simptomele bolii în studiile clinice cu medicamente *73*
Riscuri și șanse în epidemiologie *74*
Studiu epidemiologic caz-control *74*
Ratele de mortalitate, morbiditate și fertilitate *74*
Ratele de incidență în analiza de supraviețuire a cohortelor de cazuri *75*
Prevalența *75*
Rata mortalității *75*
Estimarea dimensiunilor eșantionului *76*
Referințe 77
Anexa 78

3. ANALIZA DATELOR CU PROGRAMARE R 81
Introducere 81
3.1 *Date și prelucrare a datelor 82*
Codarea datelor *83*
Captura de date *84*
Editarea datelor *84*
Imputații *85*
Calitatea datelor *85*
Producerea rezultatelor *87*
Întrebări de revizuire pentru Secțiunea 3.1 *87*
3.2 *Începând cu R 87*
R și Biostatistică *88*
O primă sesiune folosind R *90*
Mediul R *97*
Întrebări de revizuire pentru Secțiunea 3.2 *98*
3.3 *R ca calculator 99*
Operații matematice folosind R *99*
Atribuirea valorilor în R și calcule Utilizarea vectorilor și matricilor *100*
Calcule în vectori și grafice simple *101*
Utilizarea factorilor în programarea R *101*
Grafică simplă *103*
x ca vectori și matrici în biostatistică *107*
Câteva funcții speciale care creează vectori *108*
Matrice și matrice *108*
Utilizarea funcției de dimensiune dim în R *109*

Utilizarea matricei Matrix Function în R *109*
Câteva funcţii utile care funcţionează pe matrici în R *110*
NA: „Nu este disponibil” pentru valorile lipsă din seturile de date *111*
Funcţii speciale care creează vectori *111*
Întrebări de revizuire pentru Secţiunea 3.3 *112*
Exerciţii pentru Secţiunea 3.3 *114*
3.4 *Utilizarea R în analiza datelor în BIOS* *115*
Introducerea datelor la linia de comandă R *116*
de funcţii () şi Making of data.frame() în R *127*
Întrebări de revizuire pentru Secţiunea 3.4 *130*
Exerciţii pentru Secţiunea 3.4 *130*
3.5 *Analiza datelor univariate, bivariate şi multivariate* *133*
Analiza univariată a datelor *133*
Analiza datelor bivariate şi multivariate *136*
Analiza multivariată a datelor *138*
Analiza varianţei (ANOVA) *140*
Întrebări de revizuire pentru Secţiunea 3.5 *147*
Exerciţii pentru Secţiunea 3.5 *147*
Referinţe *150*
Anexă: Documentaţia pentru funcţia de plot *151*
Graficul X–Y generic *151*

4. GRAFICE CU R 155
Introducere *155*
Alegerea sistemului *156*
Pachete *156*
4.1 *Grafică de bază (sau tradiţională)* *157*
Funcţii de nivel înalt *157*
Funcţii de trasare la nivel scăzut *161*
Interacţiunea cu grafica *165*
Utilizarea parametrilor grafici *165*
Lista parametrilor pentru grafică *166*
Drivere de dispozitiv *168*
Întrebări de revizuire pentru Secţiunea 4.1 *170*
Exerciţii pentru Secţiunea 4.1 *171*
4.2 *Grafică grilă* *183*
Pachetul lattice : Trellis Graphics *183*
Modelul grilă pentru R Graphics *189*
Obiecte grafice grilă *192*
Aplicaţii la investigaţiile biostatistice şi epidemiologice *194*
Întrebări de revizuire pentru Secţiunea 4.2 *207*
Exerciţii pentru Secţiunea 4.2 *207*
Referinţe *210*

5. PROBABILITATE ŞI STATISTICĂ ÎN BIOSTATISTICĂ 211
Introducere *211*
5.1 *Teorii ale probabilităţii* *212*
Ce este probabilitatea? *212*
Proprietăţile de bază ale probabilităţii *213*
Calcule de probabilitate folosind R *223*
Aplicaţii ale teoriei probabilităţilor în ştiinţele

sănătății *227*
Statistici rezumative tipice în biostatistică: intervale de încredere, teste de semnificație și bunătatea potrivirii *232*
Întrebări de revizuire pentru Secțiunea 5.1 *257*
Exerciții pentru Secțiunea 5.1 *258*
5.2 *Inferență statistică tipică în biostatistică: biostatistică bayesiană 268*
Ce este biostatistica bayesiană? *268*
Teorema lui Bayes în teoria probabilității *268*
Metodologia bayesiană și modelele de analiză a supraviețuirii (timpul până la eveniment) pentru biostatistică în epidemiologie și medicină preventivă *268*
Formula Bayes inversă *288*
Modelarea în biostatistică *298*
Întrebări de revizuire pentru Secțiunea 5.2 *304*
Exerciții pentru Secțiunea 5.2 *305*
Referințe 321

6. STUDII DE CAZ-CONTROL ŞI STUDII DE COHORT ÎN EPIDEMIOLOGIE 323
Introducere 323
6.1 *Teoria și analiza studiilor caz-control 324*
Avantajele și limitările studiilor caz-control *324*
Analiza studiilor caz-control *325*
Întrebări de revizuire pentru Secțiunea 6.1 *327*
Exerciții pentru Secțiunea 6.1 *328*
6.2 *Teoria și analiza studiilor de cohortă 328*
O aplicație importantă a studiilor de cohortă *329*
Studii clinice *329*
Studii controlate randomizate *329*
Studii de cohortă pentru bolile la alegere și bolile netransmisibile *330*
Studii de cohortă și diagramă lexic în biostatistica demografiei *336*
Întrebări de revizuire pentru Secțiunea 6.2 *338*
Exerciții pentru Secțiunea 6.2 *339*
Referințe 344

7. ÎNCERCĂRI ALEATORIZATE, DEZVOLTARE FAZĂ, CONFUNDARE ÎN SUPRAVIEȚUVIERE ANALIZA ȘI REGRESII LOGISTICE 345
7.1 *Studii randomizate 345*
Clasificări ale RT-urilor după proiectarea studiului *345*
Randomizare *346*
Analiza biostatistică a datelor din RT *348*
Biostatistică pentru RT în mediul R *348*
Întrebări de revizuire pentru Secțiunea 7.1 *350*
Exerciții pentru Secțiunea 7.1 *350*
7.2 *Faza de dezvoltare 361*
Faza 0 sau Faza preclinica *361*
Faza I *363*
Faza II *363*
Faza III *363*
Farmacoepidemiologie: o ramură a epidemiologiei *363*

Câteva teste de bază în dezvoltarea fazei epidemiologice *364*
Întrebări de revizuire pentru Secţiunea 7.2 *369*
Exerciţii pentru Secţiunea 7.2 *369*
7.3 *Confuzie în analiza supravieţuirii* *375*
Abordări biostatistice pentru controlul confuziei *376*
Utilizarea modelării de regresie pentru controlul confuziei *385*
Confuzie şi coliniaritate *386*
Întrebări de revizuire pentru Secţiunea 7.3 *389*
Exerciţii pentru Secţiunea 7.3 *389*
7.4 *Regresii logistice* *412*
Inadecvarea regresiei liniare simple când *y* este a Variabila dependentă categorială *413*
Modelul de regresie logistică *413*
Logit *414*
Analiza regresiei logistice *414*
Modele liniare generalizate în R *415*
Întrebări de revizuire pentru Secţiunea 7.4 *419*
Exerciţii pentru Secţiunea 7.4 *420*
Referinţele 437
Index 439

Prefaţă

Un ştiri de televiziune naţională a difuzat următoarele cinci articole în 2006 [1]:

- Un raport despre utilizarea medicamentelor netradiţionale, non-mainstream, cum ar fi tradi
  medicina chineză naţională (MTC), pentru a ajuta persoanele care au diabet de tip 1
- Îngrijorarea unui cartier cu privire la o creştere bruscă a numărului de copii cu astm şi autism care trăiesc într-un stat din vestul SUA
- Un raport despre cele mai recente recomandări ale Centrelor pentru Controlul şi Prevenirea Bolilor (CDC) cu privire la cine ar trebui să primească vaccinul antigripal şi când
- O discuţie despre strategia extinsă de monitorizare a bolilor infecţioase implementată într-un oraş de coastă din sudul Statelor Unite, care a fost lovit recent de un uragan masiv
- Un raport asupra unui studiu, publicat într-un jurnal medical de top, despre o probabilă asociere între un risc crescut de cancer şi expunerea lucrătorilor la o anumită substanţă chimică

Fiecare dintre aceste ştiri a inclus interviuri cu oficiali de sănătate publică sau anchetatori care s-au autointitulat *epidemiologi* .

Cine sunt aceşti epidemiologi şi ce fac ei? Ce este *epidemiologia* ? Această carte este menită să răspundă la aceste întrebări. Procedând astfel, descrie ce este epidemiologia, cum a evoluat, cum este folosită astăzi şi care sunt unele dintre metodele şi conceptele sale cheie. Accentul se pune pe epidemiologie în practica sănătăţii publice, adică tipul de epidemiologie care se face la departamentele de sănătate locale, de stat şi naţionale.

Analiza datelor – prelucrarea informaţiilor colectate prin observare sau experimentare – este o parte foarte importantă a investigaţiilor epidemiologice. Prin urmare, stadiul tehnicii în studiile epidemiologice este avansat în mod constant pe măsură ce capacităţile şi capacitatea instalaţiilor de calcul şi mediul de calcul în general avansează. Astăzi, epidemiologii din întreaga lume pot alege dintre multe pachete de software biostatistice disponibile comercial şi utilizate pe scară largă.

Un pachet software relativ nou numit R , dezvoltat în 1993 şi disponibil gratuit prin Internet, este cel mai promiţător. R are multe funcţii avansate de modelare a regresiei, cum ar fi regresia multiliniară, regresia logistică, analiza supravieţuirii şi modelarea pe mai multe niveluri. Sprijinit aşa cum este de experţi de top în biostatistică din întreaga lume, R este acum omniprezent şi oferă tot ceea ce are nevoie un analist de date epidemiologice.

[1] S.U.A. Departamentul de Sănătate şi Servicii Umane, Centrele pentru Controlul şi Prevenirea Bolilor (CDC), Biroul pentru Forţa de Muncă şi Dezvoltarea Carierei. (2006). *Principiile epidemiologiei în practica sănătăţii publice: o introducere în epidemiologia aplicată şi biostatistică* (Curs de auto-studiu SS1000, ed. a 3-a). Atlanta, GA: Autor.

Scopul acestei cărți este de a face R ușor accesibil, la un nivel practic, tuturor viitorilor epidemiologi pentru cercetare, prelucrare a datelor și prezentare. Această carte este în esență despre învățarea R, cu accent pe aplicații în epidemiologie, sănătate publică și medicina preventivă. Pentru a folosi cât mai bine acest text, cititori

ar trebui să aibă anumite cunoștințe în utilizarea de bază a computerului. Cu R și seturile de date furnizate, utilizatorii ar trebui să poată lucra prin fiecare secțiune, învățând tehnicile de gestionare a datelor, teoriile biostatistice aferente și practica analizei datelor și prezentărilor grafice. Cartea este organizată sistematic în șapte capitole , fiecare cu un număr de secțiuni principale care acoperă spectrul de coduri R aplicabile pentru aplicații biostatistice în epidemiologie și sănătate publică.

Capitolele 1 și 2 prezintă relații de interacțiune între medicină, medicina preventivă, sănătate publică, epidemiologie și biostatistică în general, precum și concepte speciale care au fost (și sunt în curs de dezvoltare) pentru a aborda problemele cantitative din epidemiologie și sănătatea publică în special. Este prezentată o trecere în revistă a elementelor de bază din teoria probabilității pentru a introduce sau întări capacitatea cititorilor de a gestiona acest important concept de bază.

Capitolul 3 acoperă manipularea simplă a datelor folosind programarea R , în timp ce capitolul 4 prezintă capabilitățile grafice disponibile în R . În urma acestor incursiuni inițiale în R , Capitolul 5 oferă o privire de ansamblu asupra teoriei probabilității și a statisticilor matematice , ceea ce este necesar deoarece ambele domenii au devenit părți integrante ale aplicațiilor biostatistice în epidemiologie.

Capitolul 6 arată modul în care R poate fi utilizat eficient pentru a gestiona problemele clasice în studiile caz-control și investigațiile de cohortă în epidemiologie. În mod similar, analiza de supraviețuire, coloana vertebrală a multor cercetări epidemiologice, găsește un sprijin excelent în mediul R , așa cum este subliniat în Capitolul 7 .

Pentru a ajuta și a provoca cititorii, un set de „întrebări de revizuire" apare la sfârșitul fiecărei secțiuni principale. Acestea vor ajuta cititorii să-și amintească și să noteze conceptele esențiale discutate în corpul textului. Deoarece este în primul rând un subiect cantitativ, biostatistica poate fi apreciată cel mai bine prin realizarea unor exerciții adecvate, specifice și practice care implică conceptele introduse în text. Exercițiile care apar la sfârșitul majorității secțiunilor vor ghida cititorii prin aplicarea acestor idei în lumea epidemiologiei reale și a sănătății publice în cursul exersării abilităților lor în calcul folosind R . **Ghidul de studiu al studenților online îi conduce pe studenți prin soluții la exercițiile din carte și este disponibil la www.springerpub.com/ chan-biostatistica. De asemenea, pe www.springerpub.com/chan-biostatistics este un - capitol mental suplimentar intitulat Research-Level Applications of** R. **An Instructor's Man ual este, de asemenea, disponibil prin e-mail la textbook@springerpub.com.**

Pe 6 noiembrie 2011, profesorul Tomás Aragon, MD, DrPH, de la Universitatea din California-Berkeley, în prefața manualului său online *Applied Epidemiology Using R* , a făcut următorul comentariu:

> Ne place să ne gândim la R ca la un set de instrumente extensibile pentru a-și implementa planul de analiză, indiferent de simplitate sau complexitate. Speranța noastră este că tot mai mulți epidemiologi vor adopta R pentru aplicații epidemiologice, sau cel puțin îl vor include în cutia lor de instrumente.

Autorul speră că această carte va satisface această nevoie, ajutând la introducerea R , un limbaj de calcul de nivel înalt și un mediu pentru calculul biostatistic și prezentările grafice, epidemiologilor și analiștilor de date din sănătatea publică și medicina preventivă care desfășoară activ investigații epidemiologice.

Bertram KC Chan, PhD, PE

Distribuie
Biostatistică pentru epidemiologie și sănătate publică folosind R



ONE

Introduction

1.1 MEDICINA, MEDICINA PREVENTIVA, SANATATE PUBLICA SI EPIDEMIOLOGIE

Medicament

Cuvântul *medicină* este derivat din expresia latină *ars medicina* , care înseamnă „ arta vindecătoare". Astfel, *medicina* se referă la arta și știința vindecării, care folosește o varietate de practici evoluate sau dezvoltate pentru menținerea și restabilirea sănătății prin tratamentul și prevenirea bolilor și rănilor.

Medicina actuală aplică științele sănătății, cercetarea biomedicală și tehnologia - pentru a diagnostica și trata leziunile și bolile, de obicei prin medicamente/medicamente sau intervenții chirurgicale, dar și prin terapii la fel de diverse precum psihoterapia, protezele (de exemplu, membre artificiale) și kinetoterapie, pentru a numi câteva. Având în vedere această gamă largă de abordări și tehnici posibile, este firesc ca în cadrul profesiei medicale să apară diverse specialități, pe măsură ce practicienii și-au concentrat talentele și eforturile pe anumite probleme și terapii. Dezvoltarea unei specialități, inclusiv specialitatea de medicină preventivă, este adesea condusă de noua tehnologie.

Medicină preventivă și sănătate publică

În Statele Unite (precum și în Regatul Unit și în multe alte părți ale lumii), *medicina preventivă* este una dintre cele 24 de specialități medicale recunoscute de Consiliul American de Specialități Medicale (ABMS). Acesta cuprinde trei domenii de subspecializare :

1. Medicina generala preventiva si sanatate publica (PH)
2. Medicina aerospațială
3. Medicina muncii

Pentru a deveni certificat de consiliu într-una dintre domeniile de specializare în medicină preventivă , un medic licențiat din SUA trebuie să finalizeze cu succes un program de rezidențiat în medicină preventivă după un stagiu de 1 an. După aceea, medicul trebuie să finalizeze un an de practică în domeniul respectiv de specialitate și să treacă examenul comisiei de medicină preventivă. Programul de rezidențiat, care este de cel puțin 2 ani

durata, include absolvirea unei diplome de master în sănătate publică (MPH) sau echivalentă. (Prezentul text este îndreptat în primul rând către realizarea acestui ultim reper.)

De exemplu, în Statele Unite, Universitatea Loma Linda (LLU), California, oferă un program de rezidență în medicină familială și preventivă care combină pregătirea în medicina de familie și medicina preventivă, contribuind astfel la îndeplinirea misiunii LLU: „To Make Man Whole". Acest program special include formare în îngrijirea primară prin programul de rezidență în medicină de familie LLU, precum și munca în sistemele de îngrijire și îngrijire a sănătății bazate pe populație prin programul de rezidență în medicină preventivă LLU . Pe parcursul celor 4 ani în program, toți rezidenții de succes obțin o diplomă MPH prin Școala de Sănătate Publică LLU și au o oportunitate de a se expune unic la cele două domenii de forță ale LLU: medicina stilului de viață și sănătatea globală.

Sănătate Publică și Epidemiologie

Într-un studiu major realizat de Institutul de Medicină al Academiei Naționale de Științe din SUA, Comitetul pentru Studiul Viitorului Sănătății Publice a definit misiunea **sănătății publice** ca fiind „împlinirea interesului societății de a asigura condițiile în care oamenii pot fi sănătoși. " (vezi Centrele pentru Controlul Bolilor și Prevenire [CDC] 2006) . Același studiu a definit substanța sănătății publice ca „eforturi organizate ale comunității care vizează prevenirea bolilor și promovarea sănătății. [Sănătatea publică] leagă multe discipline și se bazează pe miezul științific al epidemiologiei."

Epidemiologia (EPDM) , practic, este studiul demografiei proceselor bolii, incluzând, dar fără a se limita la, studiul epidemilor. Departamentul de Sănătate și Servicii Umane din SUA (DHHS), prin CDC, oferă următoarea definiție a epidemiologiei:

> Cuvântul *epidemiologie* provine din cuvintele grecești *epi* , care înseamnă pe sau pe, *demos* , adică oameni și *logos* , adică studiul. Astfel, cuvântul *epidemiologie* își are rădăcinile în studiul a ceea ce se întâmplă cu o populație . Au fost propuse multe definiții, dar următoarea definiție surprinde principiile de bază și spiritul de sănătate publică ale epidemiologiei:
>
> Epidemiologia este studiul distribuției și factorilor determinanți ai stărilor sau evenimentelor legate de sănătate în populații specificate și aplicarea acestui studiu la controlul problemelor de sănătate. (CDC, 2006)

Întrebări de revizuire pentru Secțiunea 1.1

1. Folosind surse de pe Internet, numiți cinci specialități medicale (pe lângă medicina preventivă) din Statele Unite, care sunt recunoscute oficial de ABMS.
2. (a) Cunoașteți vreun medic care practică *numai* medicina preventivă?
   (b) Dacă ați fi medic, ați alege să practicați numai medicina preventivă ? De ce sau de ce nu?
3. *Cercetare și politică în domeniul sănătății (HRP)* : Școala de Medicină a Universității Stanford (Stanford, California) predă medicina preventivă în cadrul departamentului său de HRP. Acest program are patru domenii: biostatistică (BIOS), coordonarea datelor, EPDM și

cercetarea serviciilor de sănătate (HSR). Ultima zonă, HSR, se concentrează pe analizarea și compararea costurilor, riscurilor și beneficiilor strategiilor de îngrijire medicală, în special intervențiile medicale. Discutați posibilele beneficii ale medicinei preventive ale muncii în domeniile HRP.

4. *Fluorizarea apei potabile:* în Statele Unite, fluorizarea apei potabile a făcut obiectul multor cauze în instanță în care activiști politici au dat în judecată guvernele locale, susținând că drepturile lor la consimțământul informat pentru tratament medical și la un proces echitabil sunt încălcate prin fluorizarea obligatorie. . Oamenii au dat în judecată municipalitățile pentru boli despre care credeau că sunt cauzate de fluorizarea alimentării cu apă a unui oraș. În cele mai multe dintre aceste cazuri, instanțele au dat în favoarea orașelor pârâte sau a entităților guvernamentale, constatând nicio legătură sau doar o legătură slabă între problemele de sănătate și fluorizarea apei pe scară largă. Până în prezent, nicio instanță federală sau instanță supremă de stat nu a considerat că fluorizarea apei este ilegală. Dacă ai fi furnizor de PH, ai sprijini fluorizarea apei de băut? De ce sau de ce nu?

1.2 SĂNĂTATE PERSONALĂ ȘI SĂNĂTATE PUBLICA (SP)

Astăzi, medicina obișnuită se îndreaptă către un standard de practică bazată pe dovezi , folosind date colectate în principal din cercetarea PH și în special din sondajele epidemiologice. Aceste date dezvăluie tendințe, identifică probleme și permit evaluarea eficacității diferitelor abordări pentru anumite populații. Avantajele asistenței medicale bazate pe dovezi par evidente, prin faptul că mărește capacitatea furnizorilor de servicii medicale de a oferi tratamente eficiente și sigure pentru orice afecțiune dată.

Datele PH stau la baza recomandărilor privind sănătatea personală. Potrivit experților, toate planurile personale de sănătate pot fi rezumate simplu după cum urmează:

- Mănâncă bine și odihnește-te adecvat.
- Faceți exerciții fizice energic și regulat.

Cu toate acestea, pot apărea provocări serioase atunci când generalizările PH sunt aplicate situației personale a unui individ.

Iată un exemplu. Epidemiologia PH observațională a stilului de viață a arătat că consumul de alimente din cereale integrale este mai bine decât consumul de alimente procesate din care au fost îndepărtați mulți nutrienți, vitamine și fibre care apar în mod natural. Știind acest lucru, s-ar putea aplica aceste constatări la propria situație personală. O persoană ar putea lansa o campanie de a mânca numai pâine integrală, paste și chiar pizza 100% din grâu. Acest lucru ar putea aduce beneficii stării nutriționale a persoanei, deoarece ar consuma mult mai multe fibre și vitamine B, care sunt în general bune pentru sănătate.

Dar dacă această persoană este intolerantă la grâu și la produsele aferente?

Ce se întâmplă dacă această persoană are o problemă de intoleranță la gluten (glutenul se găsește în alimentele care conțin grâu, secară și orz)?

Ce se întâmplă dacă această persoană are o alergie la grâu?

Dacă o persoană are oricare dintre aceste afecțiuni, consumul de alimente din grâu integral îl va îmbolnăvi în mai multe moduri posibile. Reacțiile alergice pot pune viața

în pericol ; boala celiacă ar provoca probleme gastrointestinale grave; intoleranța la grâu ar putea cauza o absorbție și asimilare slabă a însăși nutrienții care au fost

menite a fi benefice. În aceste din urmă cazuri, consumul de orez din cereale integrale, de exemplu, ar putea fi cea mai bună modalitate de a obține beneficiile pentru sănătate ale unei abordări „cu cereale integrale" fără a provoca intoleranțe sau alergii alimentare.

Acesta este un exemplu simplu, dar obișnuit, în care recomandările generale de PH care ar putea aduce beneficii populației în general nu pot beneficia neapărat anumitor indivizi personal și specific. Fiecare individ ar trebui să decidă ce este potrivit pentru el sau ea atunci când creează un program individualizat eficient de îngrijire a sănătății.

Astfel de afecțiuni și probleme personale de sănătate trebuie luate în considerare ori de câte ori este propusă sau acceptată o politică de PH. Astfel, legiuitorii care au în vedere politicile de îngrijire a PH se confruntă adesea cu sarcina dificilă de a decide ce să facă cu sfatul „expert" pe baza observațiilor și concluziilor extrase din cercetările și sondajele epidemiologice. Aceștia trebuie să decidă ce pondere să acorde constatărilor și concluziilor cercetării, oportunitatea de a baza politicile publice generale pe acele constatări și utilitatea și siguranța oricăror mandate de PH la scară largă.

Ce se întâmplă dacă dovezile indică faptul că oamenii în general (sau chiar un subgrup de persoane care au un diagnostic specific, cum ar fi diabetul) se descurcă mai bine sau mai rău cu o anumită terapie? Poate fi dificil și confuz să decidă cel mai bun program: selectarea unui tratament mai familiar sau mai accesibil pentru o anumită problemă de sănătate , mai degrabă decât tratamente care sunt mai puțin familiare și mai dificil de realizat, ar putea împiedica o persoană să maximizeze beneficiile. Pentru multe persoane cu boli cronice, de exemplu, adăugarea de suplimente de vitamine, minerale și ierburi la alimente este o modalitate de a începe, dar de obicei nu este suficient. Obținerea sănătății și vindecării adevărate este o întreprindere complexă, dar plină de satisfacții. O perspectivă mai amplă poate clarifica calea prin care se ajunge la obiectivele cuiva.

Sănătatea personală versus sănătatea publică

După cum sa discutat mai devreme în această secțiune, există momente în care sănătatea personală și HP pot intra în conflict. În multe cazuri, aceasta înseamnă doar că indivizii vor face alegeri personale care se abat de la recomandările generale. Din păcate, este, de asemenea, foarte posibil ca PH să fie folosit (fie ca motiv, fie ca camuflaj) pentru a impune voința și politica statului și a trece peste drepturile cetățenilor. Inițiativele și mandatele PH sunt, în multe cazuri, determinate și administrate de sau sub o ramură a guvernului cu un obiectiv care promovează atât politicile guvernamentale, cât și agendele politice .[2] Interesele personale, aspirațiile, alegerile individuale, libertatea și idealurile pot fi sever restricționate de legile și politicile guvernamentale de „sănătate publică". Unele dintre numeroasele exemple în acest sens includ:

- Politica privind un singur copil per cuplu căsătorit în Republica Populară Chineză
- Interdicția în Statele Unite între 1920 și 1933, care a încercat

  (fără succes) să interzică alcoolul, un puternic drog psihoactiv care astăzi este ieftin și disponibil pe scară largă. Deși datele PH arată clar că daunele provocate de acest medicament sunt uriașe, legislația guvernamentală care facea ilegală utilizarea

[2] https://en.wikipedia.org/wiki/Public_health

acestuia a fost în cele din urmă răsturnată.

- Punerea în carantină obligatorie a unei persoane care suferă de o boală infecțioasă potențial fatală.
- Imunizări obligatorii care implică durere, inconveniente și risc de efecte secundare pentru întreaga populație, astfel încât o boală să poată fi prevenită într-o minoritate . Acest lucru poate fi justificabil - de exemplu, vaccinurile care au dus la eradicarea variolei și a poliomielitei în Statele Unite - dar există o zonă mare, gri din punct de vedere etic, în jurul vaccinurilor pentru care există o incertitudine considerabilă cu privire la echilibrul dintre beneficii și daune. .
- O problemă continuă în Statele Unite astăzi este îngrijorarea multor părinți că vaccinarea poate provoca autism la copii. Ar trebui să aibă acești cetățeni dreptul de a refuza vaccinarea în numele copiilor lor? Probleme similare există atunci când convingerile religioase ale unei persoane interzic transfuziile de sânge; astfel de persoane au refuzat transfuziile de sânge pentru ei înșiși și pentru copiii lor dependenți , chiar și în situații de urgență medicală care pun viața în pericol. Deși dreptul constituțional la libertatea religiei și la exercițiul religios a depășit de obicei afirmațiile guvernului în aceste cazuri, în unele cazuri autoritățile guvernamentale pot declara o persoană „secție a instanței" și, prin urmare, își pot asuma întreaga responsabilitate atât pentru problemele personale, cât și pentru sănătatea publică.

Întrebări de revizuire pentru Secțiunea 1.2

1. Unde se termină sănătatea personală și unde începe HP?
2. Diagrama Venn din figura 1.1 reprezintă cu exactitate sănătatea personală și PH? Enumerați câteva probleme care sunt:
   (a) Contează exclusiv sănătatea personală
   (b) Contează exclusiv PH-ul
   (c) Probleme atât de sănătate personală, cât și de sănătate publică
   Spuneți motivele care susțin identificarea dvs.

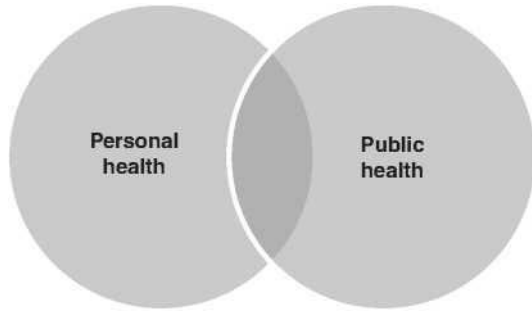


FIGURA 1.1 Sănătatea personală și PH.

1.3 CERCETĂRI ȘI MĂSURĂTORI ÎN EPDM ȘI PH

După cum sa discutat în Secțiunea 1.1, cuvântul *epidemiologie* înseamnă literal „studiul a ceea ce se întâmplă cu o populație". De-a lungul anilor, au fost propuse multe definiții,

dar următoarea definiție surprinde principiile de bază și spiritul PH al EPDM:

> Epidemiologia este **studiul** distribuției și **factorilor determinanți** ai **stărilor sau evenimentelor legate de sănătate** în **populații specificate** și **aplicarea** acestui studiu la controlul problemelor de sănătate (Broadbent, 2009).

Termenii critici din această definiție (îngroșate aici) reflectă unele dintre principiile importante ale EPDM:

1. **Studiază** . EPDM este o disciplină științifică care utilizează metode consacrate de cercetare științifică. Este bazat pe date și depinde de o abordare sistematică și imparțială a colectării, analizei și interpretării datelor. Metodele epidemiologice depind de observarea și utilizarea unor grupuri de comparație valide pentru a evalua dacă ceea ce a fost observat (cum ar fi numărul de cazuri de boală într-o anumită zonă într-o anumită perioadă de timp sau frecvența expunerii în rândul persoanelor cu boală) diferă de ceea ce ar putea fi de așteptat. EPDM utilizează, de asemenea, metode din alte domenii științifice, inclusiv BIOS și informatică și alte științe biologice, economice, sociale și comportamentale. Această carte tratează în principal contribuția specială a BIOS-ului la EPDM.
2. **Distributie** . EPDM raportează **frecvența** și **modelul** evenimentelor de sănătate și rezultate într-o populație de interes:
   - **Frecvența** se referă la numărul de evenimente de sănătate, cum ar fi numărul de cazuri de holeră sau diabet într-o populație și la relația dintre acest număr și dimensiunea populației. Rata rezultată permite epidemiologilor să compare apariția bolii la *diferite* populații.
   - **Modelul** se referă la apariția evenimentelor legate de sănătate în funcție de persoană, timp și loc. Tiparele *de timp* sunt cele găsite în anumite perioade de timp care influențează apariția rănilor sau a bolii, cum ar fi anual, sezonier, săptămânal, zilnic, orar, în ziua săptămânii versus weekend și așa mai departe. Tiparele *de locație* includ variații geografice, diferențe urbane versus rurale și locația locurilor de muncă sau școlilor. Tiparele caracteristice *personale includ* factori demografici care sunt sau pot fi legați de riscul de boală, rănire sau dizabilitate, cum ar fi vârsta, sexul, starea civilă și statutul socioeconomic, precum și comportamentele și expunerile la mediu.
3. **Determinanti** . Determinanții sunt factori – fie că sunt evenimente, caracteristici sau alte lucruri – care provoacă o schimbare a stării de sănătate sau a altor caracteristici definite. În acest domeniu de investigație, cauzele bolilor sunt îndeaproape studiate, identificate și corelate cu rezultatele așteptate și măsurate asupra sănătății. Acest aspect critic al EPDM este discutat mai detaliat mai târziu în această secțiune .
4. **Stări sau evenimente legate de sănătate.** Inițial, EPDM era preocupat doar de epidemiile de boli transmisibile (cum ar fi holera). Ulterior, totuși, domeniul de anchetă a fost extins pentru a aborda cazurile endemice (restrânse sau specifice

   la o localitate sau regiune) boli transmisibile și boli infecțioase netransmisibile. Până la mijlocul secolului al XX-lea, metode epidemiologice suplimentare au fost dezvoltate și aplicate la leziuni, boli cronice, sănătatea materno-copilă și malformații congenitale, sănătatea mediului și sănătatea ocupațională. Acum, epidemiologii investighează și comportamentele legate de sănătate și bunăstare, cum ar fi probleme legate de stilul de viață, cum ar fi cantitatea de exerciții fizice efectuate, și probleme de siguranță, cum ar fi folosirea centurii de siguranță a mașinii. Mai mult, odată cu dezvoltarea metodelor biomoleculare și secvențierea genomului uman, epidemiologii

au început să examineze markerii genetici ai bolilor și factorii de risc ale bolii. Ca urmare, termenul de evenimente sau *stări legate de sănătate* poate fi folosit în general pentru a se referi la orice problemă care afectează bunăstarea unei populații. Cu toate acestea, în EPDM modern, termenul de *boală* reprezintă gama largă de stări și evenimente legate de sănătate care sunt studiate.

5. **Populații specificate.** Chiar dacă atât medicii, cât și epidemiologii sunt preocupați de apariția și controlul bolilor, ei diferă în ceea ce privește modul în care ei văd „pacientul". Medicul și alți furnizori de servicii medicale sunt preocupați în primul rând de sănătatea unui individ; epidemiologul este preocupat de sănătatea colectivă a persoanelor care alcătuiesc o comunitate sau o populație . Astfel, medicul și epidemiologul au responsabilități diferite față de o persoană care are o boală. De exemplu, atunci când un pacient prezintă boală diareică, ambii profesioniști medicali sunt interesați să stabilească diagnosticul corect. Cu toate acestea, medicul se concentrează pe tratarea bolii și îngrijirea pacientului ca individ; epidemiologul se concentrează pe identificarea sursei sau a tipului de expunere care a cauzat boala, a numărului de alte persoane care ar fi putut fi expuse în mod similar, a potențialului de răspândire în continuare în comunitate și a intervențiilor pentru a preveni cazuri suplimentare sau recidive.
6. **Aplicație.** EPDM este implicat atât în studierea sănătății într-o populație, cât și în aplicarea cunoștințelor dobândite din aceste studii în practica comunitară . Ca și practica medicinei, practica EPDM este atât o știință, cât și o artă. Pentru a pune un diagnostic adecvat și a prescrie un tratament adecvat pentru un pacient, medicul combină cunoștințele medicale științifice bazate pe dovezi cu experiența, raționamentul clinic și înțelegerea pacientului. În mod similar , epidemiologul folosește metodele științifice ale EPDM descriptiv și analitic, precum și experiența, judecata epidemiologică și înțelegerea condițiilor locale, pentru a „diagnostica" sănătatea unei comunități și a propune intervenții PH adecvate, practice și acceptabile pentru controlul și prevenirea bolilor în comunitate.

EPDM: Știința de bază a PH

EPDM este știința de bază a PH din următoarele motive:

1. EPDM este un studiu cantitativ care se bazează pe cunoștințe de lucru despre probabilitate, BIOS și metode de cercetare științifică.
2. EPDM este o disciplină de raționament cauzal în care ipotezele din diverse domenii științifice, cum ar fi științe biologice, științe comportamentale, științe fizice și ergonomie, sunt dezvoltate și testate pentru a produce rezultate legate de sănătate.

Mai mult, EPDM nu este doar o disciplină de cercetare în sine, ci și o componentă a PH, oferind fundația și direcționând acțiunile PH adecvate și practice, bazate pe raționament științific, cauza-efect. Astfel, disciplina EPDM poate fi descrisă în termeni de două abordări: EPDM descriptiv și EPDM analitic.

EPDM DESCRIPTIV

EPDM descriptiv acoperă **timpul** , **locul** și **persoana.** Această abordare este extrem de importantă deoarece:

- La examinarea datelor, epidemiologul se familiarizează cu limitele acestora pe baza

variabilelor cunoscute. De exemplu, epidemiologii se ocupă adesea de un număr mare de înregistrări din care lipsesc date pentru fiecare variabilă importantă . Acest lucru a condus la dezvoltarea și aplicarea teoriei *analizei datelor lipsă* , care permite cercetătorilor să se ocupe de excentricitățile datelor (de exemplu, toate cazurile au vârsta cuprinsă între 3 luni și 4 ani, plus un tânăr de 19 ani).

- Epidemiologul învață amploarea și tiparul problemei HP care este studiată; de exemplu, care luni/cartiere/grupuri de persoane au cele mai multe și cele mai puține cazuri de fenomen de interes.
- Epidemiologul creează o descriere detaliată a stării de sănătate a unei populații care poate fi comunicată cu ușurință prin grafice, tabele și hărți.
- Epidemiologul poate identifica tipare (o sarcină dificilă, chiar și cu computere), cum ar fi zone sau grupuri din populație care au rate anormal de ridicate de îmbolnăvire. Aceste informații oferă, la rândul lor, indicii cu privire la cauzele bolii, care informează dezvoltarea de ipoteze verificabile și teorii aplicabile.

EPDM descriptiv poate fi rezumat în termenii „cinci W”:

- **W** hat = problemă de sănătate de îngrijorare
- **W** ho = persoană
- **W** aici = loc
- **Cand** = timp
- **De** ce/Cum = cauze, factori de risc și moduri de transmitere

EPDM ANALITICE

EPDM descriptiv observă și identifică modele între cazuri și în populații în funcție de timp, locație și persoană. Din aceste observații, cercetătorii pot dezvolta ipoteze despre cauzele acestor modele și despre factorii care cresc riscul de îmbolnăvire. Astfel, epidemiologii folosesc EPDM descriptiv pentru a genera ipoteze, dar doar rareori pentru a le testa ipotezele. Pentru acesta din urmă, epidemiologii apelează la EPDM analitic, care se caracterizează prin utilizarea grupurilor de comparație.

Ca o ilustrare (White, Armstrong și Saracci, 2008) , luați în considerare izbucnirea mare de hepatită A care a avut loc în statul Pennsylvania în 2003. Epidemiologii au descoperit că aproape toți pacienții cazați au mâncat la un anumit restaurant în timpul ultimele 2 până la 6 săptămâni (perioada tipică de incubație pentru hepatita A) înainte de debutul bolii.

Deși cercetătorii au reușit să-și concentreze atenția asupra unui singur restaurant și au reușit să excludă preparatorii și servitorii de mâncare ca sursă, ei nu știau ce mâncare anume a fost contaminată. Ei au întrebat pacienții cazului care dintre alimentele restaurantului au mâncat, dar asta a indicat doar ce alimente sunt populare. Apoi, cercetătorii au înrolat și au intervievat un grup de control: un grup de persoane care au mâncat la restaurant în aceeași perioadă, dar care *nu* s-au îmbolnăvit. Dintre cele 133 de articole din meniul restaurantului, cea mai izbitoare diferență între grupul de caz și cel de control a fost proporția care a mâncat salsa (94% dintre pacienții de caz au mâncat salsa, comparativ cu doar 39% dintre martori).

O investigație ulterioară a ingredientelor din salsa a implicat ceapa verde ca sursă de infecție. La scurt timp după aceea, Food and Drug Administration (FDA) a emis un avertisment către public cu privire la ceapa verde și riscul de hepatită A. Această acțiune

a fost ca răspuns direct la rezultatele convingătoare ale EPDM analitic, generate de compararea istoricului expunerii la caz pacienti cu cel al unui grup de comparatie adecvat.

Când se descoperă că pacienții cu o anumită caracteristică au o probabilitate mai mare decât cei fără caracteristică de a contracta o boală, se spune că acea caracteristică particulară este **asociată** cu boala. Caracteristica poate fi:

- Factorul demografic , cum ar fi locul de reședință, vârsta, rasa sau sexul
- Factori constituționali , cum ar fi grupa de sânge sau starea imunitară
- Comportament sau acțiune, cum ar fi fumatul sau consumul de salsa
- Circumstanțele, cum ar fi locuirea în apropierea solurilor contaminate sau a unui loc de deșeuri toxice sau utilizarea apei contaminate

Identificarea factorilor asociați cu o boală poate ajuta oficialii din domeniul sănătății să se concentreze pe controlul și prevenirea HP, precum și pe continuarea cercetării asupra cauzelor bolii.

Principalele funcții epidemiologice

Au fost identificate șase sarcini majore ale EPDM în practica PH:

1. Supravegherea PH
2. Investigație pe teren
3. Studii analitice
4. Evaluare
5. Legături
6. Dezvoltarea politicilor

Unele dintre aceste sarcini sunt reflectate în discuția, în Secțiunea 1.1, a concentrării departamentale HRP a Școlii de Medicină a Universității Stanford pe BIOS, coordonarea datelor, EPDM și HSR.

EPDM analitic se concentrează pe activitățile de prevenire și control al HP. De asemenea, ghidează cercetări suplimentare asupra cauzelor bolii. Astfel, EPDM analitic este preocupat de căutarea cauzelor și efectelor, sau de ce și cum. Se urmărește să cuantifice asocierea dintre expuneri și rezultate și să testeze ipotezele despre relațiile cauzale. Poate că EPDM în sine nu poate *dovedi niciodată* că un anumit

expunerea a determinat un anumit rezultat, dar EPDM oferă totuși dovezi suficiente pentru dezvoltarea și implementarea măsurilor adecvate de control și prevenire.

Studiile epidemiologice pot fi clasificate în două categorii: experimentale și observaționale.

STUDII EXPERIMENTALE EPDM

Într-un studiu *experimental* , investigatorul determină printr-un proces controlat expunerea pentru fiecare individ (într-un studiu clinic) sau comunitate (într-un studiu comunitar), apoi urmărește indivizii sau comunitățile în timp pentru a detecta efectele continue ale expunerii. . Următoarele sunt exemple de studii experimentale tipice:

1. Într-un studiu clinic al unui nou vaccin, cercetătorul desemnează aleatoriu unii dintre participanți să primească noul vaccin, în timp ce alții primesc un plas cebo. (Un

placebo este o substanță inertă sau inofensivă, utilizată în special în experimentele controlate care testează eficacitatea unei alte substanțe ca tratament.) Apoi, cercetătorul urmărește toți participanții, observând cine dezvoltă boala pe care noul vaccin este destinat să o prevină și compară cele două grupuri (noul vaccin versus placebo) pentru a vedea dacă grupul cu vaccin are o rată mai mică de boală.

2. Într-un studiu pentru a preveni apariția diabetului în rândul persoanelor cu risc ridicat, cercetătorii au repartizat aleatoriu subiecții într-unul din trei grupuri:
   - Un medicament antidiabetic
   - Un placebo
   - Intervenție în stilul de viață

   La sfârșitul perioadei de probă, cercetătorii caută cea mai scăzută incidență a diabetului și constată că cea mai mică incidență a avut loc în grupul de intervenție în stilul de viață, următoarea incidență cea mai mică a fost în grupul cu medicamente antidiabetice și cea mai mare incidență a avut loc în grupul placebo. grup

STUDII OBSERVATIONALE EPDM

În studiile observaționale EPDM, cercetătorul observă doar expunerea și starea de boală a fiecărui participant la studiu. Exemplul clasic de studiu observațional este investigația dr. John Snow asupra unei epidemii de holeră din 1854 la Londra.

POVESTEA HOLEREI.[3] O boală transmisă prin apă cunoscută sub numele de holeră s-a dovedit a fi una dintre cele mai virulente ucigașe din istorie. Prin investigarea epidemilor de holeră, epidemiologii au descoperit legătura dintre salubritate și HP – o descoperire care a condus la dezvoltarea sistemelor moderne de apă și canalizare din lume.

Acum se știe că holera este cauzată de ingerarea de apă, alimente sau alte materiale contaminate cu fecalele unui pacient sau gazdă holeră. De exemplu, contact casual

cu o oală contaminată, îmbrăcăminte sau lenjerie de pat murdară sau chiar o mână nespălată ar putea fi tot ceea ce este necesar pentru a contracta holera. Boala se transmite usor si actioneaza rapid. Lovitură atât de brusc încât o persoană poate fi sănătoasă dimineața și poate muri până seara. De la debutul simptomelor - diaree, crampe musculare, vărsături și febră - moartea poate apărea în decurs de 48 de ore sau mai puțin. Se pierde atât de mult lichid încât sângele pare gros, iar aproximativ 50% dintre pacienți vor muri, în principal din cauza deshidratării. În diverse părți ale lumii (inclusiv Europa și Asia), zeci de mii au murit din cauza acestei boli epidemice.

PĂRINTELE CÂMPULUI EPDM. La mijlocul anilor 1800, un medic anestezist pe nume Dr. John Snow a efectuat o serie de studii la Londra care l-au făcut să fie numit „părintele epidemiologiei de teren”. Dr. Snow a investigat focarele de holeră, atât pentru a descoperi cauza bolii, cât și pentru a preveni reapariția acesteia. Lucrarea sa ilustrează secvența clasică de la EPDM descriptiv la generarea de ipoteze la testarea ipotezelor (EPDM analitic) la aplicare:

[3] Informații biografice despre Dr. John Snow, „părintele” domeniului EPDM și munca sa sunt disponibile la http://www.ph.ucla.edu/epi/snow.html

- În 1854, dr. Snow a efectuat unul dintre studiile sale, când a început brusc o epidemie de holeră în Piața de Aur din Londra. El a început prin a stabili unde, în această zonă anume, trăiau și lucrau persoanele cu holeră. El a marcat fiecare reședință pe o hartă a zonei, așa cum se arată în Figura 1.2 . Acest tip de hartă, care arată distribuția geografică a cazurilor, se numește *hartă spot* .
- Deoarece dr. Snow credea că apa este o sursă purtătoare de infecție pentru colera , el a marcat și locația pompelor de apă pe harta locală și apoi a căutat o relație între distribuția gospodăriilor cu cazuri de holeră și localizarea pompele de apă. El a observat că *în jurul pompei A , pompa Broad Street, mai multe case de locuințe s-au grupat* decât în jurul pompei B sau C. Când i-a interogat pe locuitorii care locuiau în zona Golden Square, i s-a spus că au evitat pompa B pentru că era extrem de contaminată, și că Pompa C era amplasată prea incomod pentru majoritatea dintre ei.
- Din aceste informații, Dr. Snow a concluzionat că pompa Broad Street (Pompa A) a fost sursa primară de apă și cea mai probabilă sursă de infecție pentru majoritatea persoanelor cu holeră din zona Golden Square.
- El a remarcat, de asemenea, că nu a avut loc niciun caz de holeră într-o zonă de două blocuri chiar la est de pompa de pe Broad Street. În urma cercetărilor, Snow a găsit o fabrică de bere situată acolo, cu o fântână adâncă în incintă. Lucrătorii de la berărie și-au obținut apa din această fântână și au primit, de asemenea, o porție zilnică de lichior de malț. Accesul la aceste rații necontaminate (apă și lichior) ar putea explica de ce niciunul dintre angajații fabricii de bere nu a contractat holera.
- Pentru a confirma că pompa de pe Broad Street a fost sursa epidemiei , dr. Snow a adunat informații despre locul în care persoanele cu holeră și-au obținut aprovizionarea cu apă. *Consumul de apă de la pompa Broad Street a fost singurul factor comun printre bolnavii de holeră.*
- După ce dr. Snow și-a prezentat concluziile oficialilor municipali , mânerul pompei a fost îndepărtat – iar focarul s-a încheiat! (Locul pompei este acum marcat de o placă montată pe peretele din afara Pubului John Snow, numit corespunzător.)

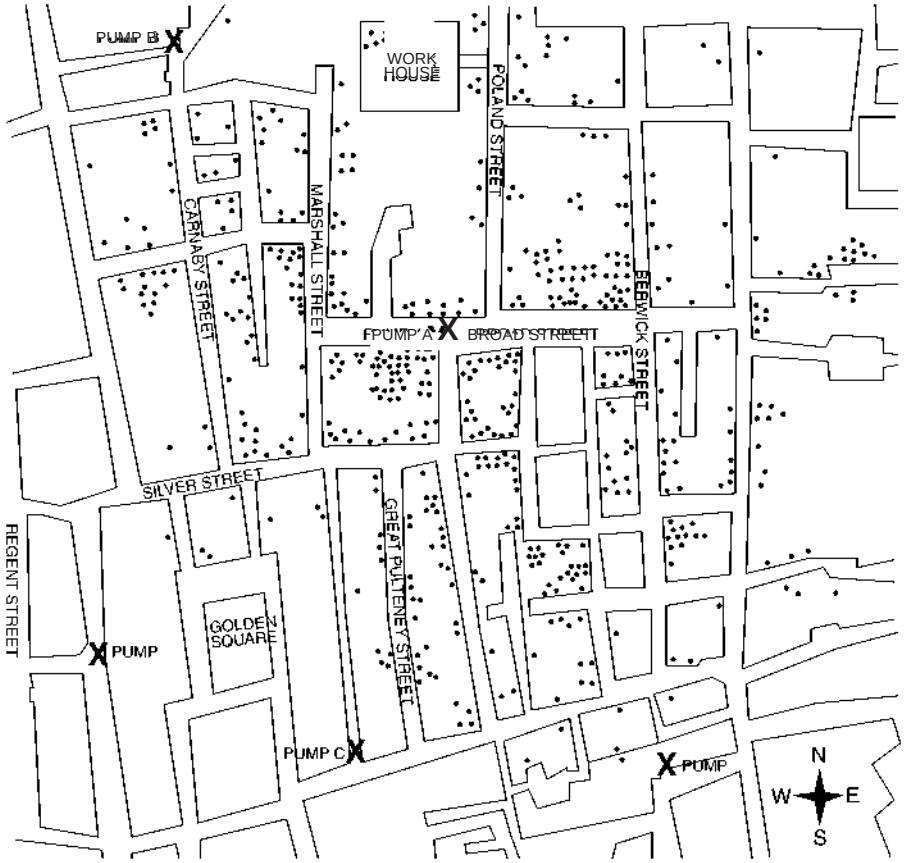


FIGURE 1.2 Spot map of deaths from cholera in the Golden Square area, London, 1854 (redrawn from original).

Source: Humphrey Milford, *Dr. John Snow on Cholera*. London: Oxford University Press, 1936.

TIPURI DE STUDII OBSERVATIONALE EPDM

Cele mai comune două tipuri de studii observaționale sunt studiile de cohortă și studiile caz-control; al treilea tip este studiile transversale.

STUDII DE COORTA. Într-un *studiu de cohortă* , indiferent dacă fiecare participant la studiu este expus sau nu:

1. Epidemiologul **înregistrează** și apoi **urmărește** fiecare participant pentru a vedea dacă el sau ea dezvoltă boala de interes. (Acest lucru diferă de un studiu experimental deoarece, într-un studiu de cohortă, epidemiologul observă mai degrabă decât determină starea de expunere a participantului.)
2. După o perioadă de timp, epidemiologul **compară** rata bolii în grupul expus cu rata bolii în grupul neexpus.
3. Grupul neexpus servește ca grup de comparație, oferind o estimare a liniei de bază sau a cantității așteptate de apariție a bolii în comunitate.
4. Dacă rata bolii este substanțial diferită în grupul expus comparativ cu grupul neexpus, se spune că expunerea este asociată cu boală.

Durata urmăririi variază considerabil. Pentru a răspunde rapid la o problemă de HP, cum ar fi un focar de holeră sau gripă aviară, departamentele de HP tind să efectueze

studii relativ scurte. Organizațiile de cercetare și academice au mai multe șanse să efectueze studii despre cancer, boli cardiovasculare și alte boli cronice, în eforturi care pot dura ani și chiar decenii. De exemplu:

- **Framingham Health Study** este un studiu de cohortă care a urmărit peste 5.000 de locuitori din Framingham, Massachusetts, de la începutul anilor 1950, pentru a stabili factorii de risc pentru bolile cardiovasculare.
- **Nurses Health Study** și **Nurses Health Study II** sunt studii de cohortă înființate în 1976 și, respectiv, 1989, care au urmărit fiecare mai mult de 100.000 de asistente și au furnizat informații utile despre contraceptivele orale , dieta și factorii de risc pentru stilul de viață.

Studiile Adventiste de Sănătate: Studiul 1. **Studiul Adventist de Sănătate 1** (AHS-1),[4] o investigație de cohortă care a început în 1974, a avut unele diferențe foarte elementare față de studiile anterioare privind mortalitatea. A fost conceput pentru a elucida care componente ale stilului de viață adventist de ziua a șaptea (SDA) oferă protecție împotriva bolilor. Acest studiu a comparat ratele de boală sau mortalitate între SDA și non-SDA. De asemenea, s-au colectat date privind evenimentele bolii nefatale, precum și fatale. Acest studiu a adăugat și o investigație mai detaliată a dietei. La început, AHS-1 a fost în primul rând o investigație a cancerului. În 1981, a fost adăugată o componentă cardiovasculară. Rata de rentabilitate pentru chestionarele anuale de urmărire SDA, care au întrebat despre spitalizări și au fost esențiale pentru întregul proces de cercetare, a fost de peste 90% și de obicei peste 95%. Trimiterea finală și cea mai critică a înregistrat un răspuns incredibil de 99,5%.

Studiile Adventiste de Sănătate: Studiul 2. [3] Studiul actual, care a început în 2002 și a stabilit un obiectiv de participare a 125.000 de ASD, continuă să exploreze legăturile dintre stilul de viață, dietă și boală în rândul bazei mai largi de adventiști din Statele Unite și Canada. În mai 2006, Studiul Adventist de Sănătate-2 (AHS-2) a avut o înscriere de 96.741 de persoane. Dr. Gary E. Fraser, împreună cu o echipă de cercetători de la Școala de Sănătate Publică de la LLU, efectuează studiul, care este finanțat de Institutul Național al Cancerului din SUA. În iulie 2011, National Institutes of Health (NIH) a acordat AHS-2 un grant substanțial pe 5 ani pentru a continua studiul.

Tipuri de studii de cohortă. Studiile AHS-1 și AHS-2 sunt uneori numite studii de cohortă **de urmărire** sau **prospective** , deoarece participanții sunt înscriși atunci când începe studiul și sunt apoi urmăriți prospectiv în timp pentru a identifica apariția rezultatelor de interes.

Un tip alternativ este un studiu de cohortă **retrospectiv** . În acest tip de studiu, atât expunerea, cât și rezultatele au avut loc deja. La fel ca într-un studiu de cohortă prospectiv, anchetatorul calculează și compară ratele bolii în grupurile expuse și neexpuse. Studiile de cohortă retrospective sunt utilizate în mod obișnuit

în investigațiile bolilor în grupuri de persoane ușor de identificat, cum ar fi muncitorii unei anumite fabrici sau participanții la o nuntă. De exemplu, un studiu de cohortă retrospectiv a fost folosit pentru a determina sursa de infecție a ciclosporiazei, o boală

[4] AHS-1 și AHS-2 sunt disponibile la http://publichealth.llu.edu și http://www.llu.edu/public- sănătate/sănătate/index.page

parazitară care a izbucnit printre membrii unei unități rezidențiale din Pennsylvania în 2004. Investigația a implicat consumul de mazăre de zăpadă ca vehicul al focarului de ciclosporiază.

Alte tipuri de studii includ studii caz-control și studii transversale.

STUDII DE CAZ-CONTROL. Într-un studiu caz-control, anchetatorii încep prin a înrola un grup de persoane cu boală (uneori numiți *pacienți caz* mai degrabă decât *cazuri* , deoarece *cazul* se referă la o apariție a bolii, nu la o persoană). Ca grup de comparație, investigatorul înscrie apoi un grup de persoane fără boală (*martori*).

Investigatorii compară apoi expunerile anterioare între cele două grupuri. Grupul de control oferă o estimare a valorii de referință sau a cantității așteptate de expunere în acea populație. Dacă cantitatea de expunere în cadrul grupului de cazuri este substanțial mai mare decât cantitatea la care s-ar aștepta pe baza grupului de control, atunci se spune că boala este asociată cu acea expunere.

Cheia într-un studiu caz-control este identificarea unui grup de control adecvat – unul care este comparabil cu grupul de caz în majoritatea privințelor – pentru a oferi o estimare rezonabilă a expunerii de bază sau așteptate.

STUDII TRANSVERSALE. În studiul observațional de tip transversal, un eșantion de persoane dintr-o populație este înrolat și expunerea acestora și rezultatele asupra sănătății sunt măsurate simultan. Studiul transversal tinde să evalueze prezența rezultatului de sănătate la un anumit moment în timp, fără a ține cont de durata.

De exemplu, într-un studiu transversal al diabetului, unii dintre cei înscriși cu diabet ar putea să fi trăit cu diabetul lor de mulți ani, în timp ce alții ar fi fost diagnosticați recent. Din punct de vedere analitic, studiul transversal este mai slab decât un studiu de cohortă sau un studiu caz-control, deoarece un studiu transversal nu poate separa, de obicei, factorii de risc pentru apariția bolii (incidență) de factorii de risc pentru supraviețuirea bolii.

Cauza Bolilor

„Care este cauza acestei boli?" nu este o întrebare ușor de răspuns. În ceea ce privește cauza bolilor, epidemiologii sunt, în general, de părere că „ **Natura** încarcă arma, dar **hrănirea** trage trăgaciul!" (*Hurarea* este suma factorilor de mediu care influențează trăsăturile și comportamentul exprimat de un organism.)

În acest moment, în medicina de masă, EPDM se confruntă cu cel puțin două întrebări critice cu privire la cauzalitatea bolii (Broadbent, 2009) :

1. Cum ar trebui să gestioneze EPDM anumite *boli* care par a fi etiologic mai complexe decât infecțiile și deficiențele pe care EPDM le-a tratat în mod tradițional?

   În prezent, bolile cronice netransmisibile (BCNC) reprezintă o proporție mai mare de decese, cel puțin în lumea industrializată, decât în urmă cu 100 de ani și atrag mai multă atenție epidemiologică. Cu toate acestea, aceste boli o fac

   nu par susceptibile de a fi definite în termenii unui singur agent cauzal. Cu alte cuvinte, **etiologia lor** este complexă. (*Etiologia* este o ramură a științei medicale care se ocupă de cauzele și originile bolilor.)
2. Cum ar trebui să răspundă EPDM la *cauzele nou identificate* ale bolii?

Deși EPDM continuă să descopere cauze din ce în ce mai complexe și surprinzătoare de mediu ale bolii, domeniul trebuie să se ocupe acum de noua categorie de cauze: genetica. Profunzimea și complexitatea cunoștințelor necesare pentru a trata atât determinanții genetici, cât și de mediu ai sănătății pun presiune asupra aspectelor cadrului conceptual al EPDM în ceea ce privește cauzalitatea bolii.

MODELE DE CAUZARE

Un model cauzal științific bun poate fi rezumat după cum urmează:

Cauza necesară a bolii D este evenimentul E dacă și numai dacă:

(i) Un eveniment E este o cauză a fiecărui caz de D;
(ii) Având în vedere anumite circumstanțe, un eveniment E nu este o cauză a niciunui eveniment non-D (adică, alte boli sau sănătate bună).

de vedere istoric, au fost propuse și utilizate mai multe modele de boală: modelul monocazual , modelul multifactorial și modelul contrastiv sunt cele primare.

Modelul monocauzal. Acest model spune că fiecare boală are o singură cauză care este necesară și uneori suficientă. Acest model este potrivit pentru bolile infecțioase precum tuberculoza (TB) și holera, împreună cu infestările parazitare și bolile de deficiență.

Cu toate acestea, este impropriu pentru bolile cu boli cronice cronice precum cancerul pulmonar sau diabetul. Este posibil ca diabetul să aibă o singură cauză necesară și, în anumite circumstanțe, suficientă, care nu a fost încă descoperită. Dar este și o posibilitate teoretică că nu există niciun motiv pentru ca diabetul să satisfacă această descriere. Și chiar dacă există, nu este clar cât de mult insistarea că *trebuie* să existe o astfel de cauză ajută la atingerea HP sau a oricăror obiective clinice, dacă nu se știe care este acea cauză. Ceea ce am putut identifica până acum sunt doar factori de risc cauzali, iar aceștia nu sunt nici necesari, nici suficienți . Astfel, pot fi ridicate obiecții importante cu privire la modelul monocauzal.

Modelul Multifactorial. Acest model domină acum EPDM, dar nici aceasta nu este o situație complet satisfăcătoare, deoarece modelul multifactorial nu reușește să recunoască ceea ce pare o diferență etiologică reală între boli precum colera și afecțiuni precum cancerul pulmonar. Modelul monocauzal a avut câteva succese izbitoare în istoria EPDM, iar aceste succese sunt lăsate neexplicate de simpla afirmație că cauzalitatea bolii este multifactorială. Dacă nu se poate explica succesele modelului monocauzal în termenii gândirii multifactoriale moderne, această abordare este la fel de nesatisfăcătoare.

Modelul Contrastiv. Acest model este defensabil pe motiv că leagă noțiunile de boală și de explicație generală, evitând în același timp naivitățile filozofice și dificultățile practice ale modelului monocauzal. Pentru ca persoana *p* să aibă boala *D* , este necesar ca:

SIMPTOME: *p* suferă de unele dintr-un set de simptome de sănătate *S* , care sunt diferențe între *p* și o clasă de contrast *X*.

CAUZE: Printre cauzele simptomelor lui *p* se numără evenimente de tipul *C1* , …, *Cn* , dintre care cel puțin unele nu sunt cauze ale absenței simptomelor *S* de la fiecare membru al *lui X*.

Ipotezele modelului contrastiv sunt următoarele:

1. Pentru a avea o boală, *p* trebuie să aibă unele simptome de sănătate precară. Aceste simptome sunt considerate parte a definiției bolii. Nu *toate* simptomele asociate cu acea boală anume trebuie să fie prezente, dar *p* trebuie să aibă *cel puțin* unul.

   Un *simptom* de sănătate proastă este o diferență observabilă între subiectul cazului și o clasă de contrast, care este doar un anumit set de oameni, dintre care unii pot fi doar ipotetici. Clasa de contrast nu trebuie să fie unică (adică aceeași pentru toată lumea). Clasa de contrast pentru un bărbat de 59 de ani poate include câțiva membri cheli, în timp ce clasa de contrast relevantă pentru un copil de 6 ani s-ar putea să nu. Acest lucru permite analizei să acopere bolile care sunt specifice vârstei, sexului și altor caracteristici (inclusiv apariția unei alte boli).
2. A avea o boală necesită ca simptomele lui *p* să fie cauzate de o anumită cauză sau cauze, care nu trebuie să fie cauze ale absenței simptomelor din clasa de contrast. Aceste cauze fac, de asemenea, parte din definiția bolii. De exemplu, pentru a avea holeră, trebuie să prezentați unele simptome de sănătate precară pe care o anumită clasă de contrast nu le are (de exemplu, diaree); iar acele simptome trebuie să fie cauzate de o anumită cauză specificată [adică prezența activă a *Vibrio cholerae* (*V. cholerae*) în intestinul subțire al lui *p*].

NOTĂ: A avea cauze nu este o cerință epistemologică: nu trebuie să știi despre *V. cholerae* pentru a considera holera drept boală. Mai degrabă, considerând holera drept boală, cineva *se angajează* să existe ceva ce satisface definiția/cerințele CAUZELOR.

Următorul pas este să aflați care este acea cauză, făcând astfel metoda modelului utilă din punct de vedere biologic.

Unii anchetatori iau în considerare, de asemenea, conceptul de *boală* , sau o simplă stare de sănătate sau o sănătate precară care nu se califică drept boală. Pentru o boală, o cauză sau cauzele anumitor simptome sunt specificate, în timp ce pentru o boală, acestea nu sunt specificate.

UN EXEMPLU PENTRU MODELUL CONTRASTIV

Un exemplu recent în care modelul contrastiv ar fi putut fi util este descoperirea rolului bacteriei *Helicobacter pylori* (*H. pylori*) în ulcerele duodenale. Această descoperire a adus tensiuni între modurile monocauzale și multifactoriale de gândire într-un contrast ridicat.

Multe discuții despre ulcer presupun că, de la descoperirea și implicarea *H. pylori* în formarea ulcerului, atât acidul *, cât și* stresul sau factorii psihosomatici au devenit irelevanți din punct de vedere etiologic. Unii au considerat pur și simplu *H. pylori* „cauza ulcerului”, deși tratamentele epidemiologice mai bine informate, cum ar fi un raport al NIH, au menținut o poziție diferită. Criticii au susținut că „accentul NIH pe factori multipli în patogeneză reflectă măsura în care multicauzalitatea este un element de bază al discursului biomedical și epidemiologic”.

Cu toate acestea, realitatea etiologică favorizează clar tratamentul multifactorial . *H. pylori* nu este nici necesară, nici suficientă pentru ulcerul duodenal și nici eliminarea acestuia de la un pacient nu este necesară sau suficientă pentru vindecarea unui ulcer. De fapt, modelul de gândire monocauzală devine o iluzie, o consecință a *dorinței* de tratamente care funcționează pe fiecare caz de boală. Din păcate, dorința nu este un ghid bun către realitate. În acest caz, multifactorialismul nu are resursele necesare pentru a exprima care ar putea fi realitatea etiologică, iar modelul monocauzal este pur și simplu

incorect. În schimb, modelul contrastiv este util aici.

În cadrul modelului contrastiv, ulcerele duodenale satisfac *simptomele* , iar *H. pylori* poate fi făcut să satisfacă *cauzele* . Cazurile în care ulcerul duodenal este prezent fără *H. pylori* pot fi tratate în unul din două moduri: fie sunt cazuri ale unei boli diferite cu aceleași simptome, fie se poate defini boala în termeni de *H. pylori* și o altă cauză, cum ar fi ca acid clorhidric excesiv în stomac. Rețineți că infecția *cu H. pylori* apare fără *simptome* în multe cazuri. În modelul contrastiv, aceste cazuri sunt o invitație la investigații suplimentare. Astfel, se poate adăuga precizie afirmației că *H. pylori* provoacă ulcer gastric și recunoaște importanța descoperirii prin reclasificarea unor cazuri de ulcer gastric ca o boală distinctă.

Trebuie remarcat faptul că modelul contrastiv lasă o componentă crucială nespecificată : nu indică în mod direct nimic despre clasa de contrast, și mai ales despre conceptul de sănătate.

POVESTEA LEBEDEI NEGRE

Vom lăsa pentru un moment discuția despre conceptul și definiția cauzei bolilor pentru a lua în considerare un incident istoric interesant privind definițiile biologice.

„Lebăda neagră" a fost o expresie în Anglia secolului al XVI-lea ca o declarație comună a imposibilității. Derive din prezumția Lumii Vechi că toate lebedele trebuie să fie albe, deoarece toate înregistrările istorice ale lebedelor au raportat că aveau pene albe; prin urmare, toate lebedele sunt, *prin definiție* , albe! Având în vedere acest context, o lebădă neagră era imposibilă, sau cel puțin inexistentă.

În 1697, exploratorul olandez Willem de Vlamingh a descoperit lebede negre (Figura 1.3) pe râul Swan din Australia de Vest. Ulterior, termenul a ajuns să se refere la o imposibilitate percepută care ar putea fi ulterior infirmată. Filosoful din secolul al XIX-lea John Stuart Mill a folosit eroarea logică *a lebădăi negre* ca un termen nou pentru a identifica falsificarea. În investigațiile EPDM, o „lebădă neagră" este un eveniment cu următoarele trei atribute:

1. În primul rând, este un *outlier* (a se vedea discuția despre probabilitate în secțiunea 2.3 a capitolului 2), deoarece se află în afara domeniului așteptărilor obișnuite, deoarece nimic din trecut nu indică posibilitatea acesteia.
2. Are un impact extrem.
3. În ciuda statutului său aberan inițial, se pot da motive plauzibile pentru apariția sa după fapt.



FIGURA 1.3 O lebădă neagră (*Cygnus atratus*), care a rămas nedocumentată în Occident până în secolul al XVIII-lea.

NATURA VERSUS NUTRIREA ÎN EPDM: ÎNVERSIREA DIABETULUI DE TIP 1? (EDELMAN, OLSEN, DUDLEY, HARRIS și ODDONE, 2004)

TESTUL HEMOGLOBINĂ A1c (HBA1c) PENTRU DIABET MELLITUS .[5] Hemoglobina este o substanță din globulele roșii care transportă oxigenul în întregul corp. La o persoană cu diabet zaharat slab controlat, zahărul se acumulează în sânge, fie pentru că organismul persoanei nu produce suficientă insulină pentru a transfera zahărul în celule, fie pentru că rezistența la insulină împiedică acest transfer. De aceea, zahărul din sânge se combină cu hemoglobina: hemoglobina devine „glicată". HbA1c, sau pur și simplu A1c, este fracțiunea principală a hemoglobinei glicozilate (glicohemoglobină; adică hemoglobina) de care se leagă glucoza. Glucoza rămâne legată de hemoglobină pe toată durata de viață a globulelor roșii (în mod normal, aproximativ 120 de zile), astfel încât nivelul HbA1c reflectă *nivelul mediu de glucoză din sânge din ultimele 4 luni* . Astfel, HbA1c este testată pentru a monitoriza controlul pe termen lung al diabetului zaharat.

- normal de HbAlc este mai mic de 7%. Diabeticii ating rareori astfel de niveluri, dar un control strict urmărește să se apropie de el.
- Nivelurile de peste 9% arată un control slab.
- Nivelurile de peste 12% arată un control foarte slab.

Se recomandă de obicei ca HbA1c să fie măsurat la fiecare 3 până la 6 luni la diabetici.

Studiul Diabetes Control and Complications Trial (DCCT; Nathan, 2014) a arătat că diabeticii care își mențin nivelurile de HbA1c aproape de 7% au șanse mult mai mari de a întârzia sau de a preveni complicațiile diabetului care afectează ochii, rinichii și nervii decât persoanele cu niveluri la 8% sau mai mult. O schimbare a tratamentului este aproape întotdeauna necesară dacă nivelul este peste 8%. Scăderea nivelului de HbA1c cu orice valoare îmbunătățește șansele unei persoane de a rămâne sănătos.

DIABET DE TIP 1. În diabetul de tip 1, pancreasul pacientului nu mai produce insulina de care persoana are nevoie pentru a supraviețui, așa că pacientul trebuie să înlocuiască insulina lipsă din alte surse (externe). Acesta este motivul pentru care diabetul de tip 1 este cunoscut și ca *diabet insulino-dependent* ; deoarece afecțiunea apare în primul rând la copii, a fost cunoscută și sub numele de *diabet juvenil* (de fapt, o denumire greșită, deoarece cazurile de debut la adulți nu sunt nemaiauzite).

Un diagnostic de diabet de tip 1 la un copil poate fi copleșitor la început. Dintr -o dată, părinții și copilul afectat trebuie să învețe cum să administreze injecții de insulină, să numere carbohidrații și să monitorizeze glicemia – iar copilul trebuie să facă acest

[5] Testul HbA1c pentru diabet; preluat de pe http://diabetes.webmd.com/guide/glycated- testul-hemoglobină-hba1c

lucru pentru viață. Deși diabetul la copii necesită îngrijire consecventă, progresele în monitorizarea zahărului din sânge și administrarea de insulină au îmbunătățit gestionarea zilnică a acestei afecțiuni.

SUBIECTUL CAZULUI: UN COPIL CU DIABET DE TIP 1. Subiectul cazului a fost un copil de 14 ani care a fost diagnosticat clinic ca suferă de diabet de tip 1 cu aproximativ 2 ani în urmă. Subiectul a fost înscris într-un test în care copilul a luat pe cale orală un - medicament prescris pe o perioadă de 3 luni. În mod interesant, medicamentul a fost o formulă de origine pe bază de plante medicinale tradiționale chineze (TCM; Liu și colab., 2014) .

În această perioadă, au fost efectuate teste de sânge A1c pentru a monitoriza progresul subiectului. Rezultatele testului A1c progresiv au fost următoarele:

$$9+ \rightarrow 8{,}4 \rightarrow 7{,}8 \rightarrow 7{,}45 \rightarrow 6{,}7\ (\%)$$

Întrebare: Tratamentul TCM a influențat pancreasul subiectului pentru a reporni producția de celule beta care produc insulină? (Insulina procesează și controlează nivelul glucozei din sânge.)

Următoarele teste pot fi luate în considerare:

- Poate că ar trebui măsurate modificări concomitente ale nivelului celulelor beta ale subiectului cazului; un astfel de test ar putea aduce mai multă lumină asupra subiectului.
- Poate că subiectul cazului ar putea fi scos din medicamentele prescrise pentru TCM, iar nivelurile de A1c să fie verificate îndeaproape ulterior pentru a vedea dacă tendința este inversată sau reversibilă.

Cum afectează acest rezultat poziția medicală acceptată conform căreia diabetul de tip 1 este permanent ireversibil? Cercetarea EPDM poate ajuta? În mod clar, în această situație este necesară o mare investigație EPDM .

De fapt, un studiu clinic în care același tratament TCM a fost administrat la mai mult de 10.000 de subiecți de caz a dus la un răspuns pozitiv (și anume, stabilitatea îmbunătățită a controlului glicemiei fără insulină) la aproximativ 30% din populația testată. Astfel de rezultate sunt o justificare puternică pentru investigații ulterioare EPDM în acest domeniu!

Măsurarea expunerii în epidemiologie

Studiile epidemiologice în cercetarea PH relaționează expunerea la agenții cauzali cu apariția unei anumite boli. Un studiu poate să nu explice pe deplin cum a apărut boala, dar, în general, înregistrează în ce circumstanțe ne putem aștepta să apară boala. Măsurarea exactă a expunerii la cauzele *presupuse* ale unei boli (adică *cauzele acceptate în mod obișnuit* ale bolii) este esențială pentru validitatea epidemiologică.

cercetare. Există teorii, principii și tehnici care pot fi aplicate pentru măsurarea unei game largi de expuneri, inclusiv factori științifici, medicali, genetici, demografici, comportamentali, psihologici, sociologici și de mediu.

Tehnicile în cercetarea epidemiologică includ:

- Utilizarea chestionarelor (deseori concepute de cercetător)
- Interviuri personale

- Extragerea informațiilor din dosarele medicale
- Utilizarea respondenților proxy
- de măsurători biologice și de mediu

Cercetarea poate include unul sau mai multe dintre următoarele:

1. O prezentare cuprinzătoare a erorii de măsurare și estimarea efectelor acesteia
2. Proiectarea, analiza și interpretarea validității și fiabilității studiilor
3. Modalitățile în care validitatea măsurătorilor poate fi îmbunătățită
4. Tehnici de maximizare a participării subiecților la studiile viitoare
5. Dezvăluirea problemelor etice relevante pentru măsurarea expunerii
6. Câteva îndrumări mai mult sau mai puțin cuprinzătoare pentru minimizarea erorii de măsurare

Măsurarea expunerii utilizează astfel metodele și abordările de control al calității pentru cele mai utilizate metode de colectare a datelor în EPDM.

Probleme suplimentare

Pentru a obține participarea maximă a subiecților relevanți la o cercetare epidemiologică, trebuie utilizate tehnici speciale. De asemenea, ar trebui luate în considerare aspectele etice inerente în măsurarea expunerii; în acest sens, sunt importante următoarele aspecte:

1. În studiile de fiabilitate și validitate care înregistrează gradul de eroare de măsurare pentru o anumită expunere, trebuie stabilite metodele de proiectare, interpretare și analiză a datelor colectate. Acest lucru este esențial deoarece astfel de studii de sprijin sunt necesare pentru a înțelege efectele măsurătorilor expunerii asupra studiului epidemiologic general.
2. Ar trebui alese metode pentru a maximiza ratele de răspuns. În acest fel, distorsiunea de selecție poate fi redusă. O astfel de abordare este esențială pentru succesul fazei de colectare a datelor a studiului.
3. Ar trebui luate în considerare aspectele etice în efectuarea cercetării epidemiologice în ansamblu ; beneficiile vor fi similare cu cele descrise la punctele 1 și 2.

Întrebări de revizuire pentru Secțiunea 1.3

1. Care sunt cele „cinci W” ale EPDM descriptiv?
2. În investigația de succes observațională EPDM a Dr. John Snow asupra holerei,
   (a) Ce presupuneri de bază a făcut el care l-au determinat să ajungă la concluzia corectă cu privire la cauza holerei?
   (b) Dr. John Snow s-a bazat pe un model pentru boli? Care model?

3. Din povestea lebedelor negre, este rezonabil să sugerăm că ar putea exista și lebede roșii/portocalii/galbene/verzi/albastre/indigo/violet? De ce sau de ce nu?
4. (a) Având în vedere cele trei modele pentru boli, pot exista zone suprapuse în care să se potrivească oricare două sau toate trei modele?
   (b) Desenați o diagramă Venn pentru a arăta posibilitatea (dacă există) de suprapunere a acestor trei modele. Propune câteva exemple.
5. (a) Ce este un studiu de sănătate de cohortă de caz în EPDM?
   (b) Numiți două studii de sănătate din SUA care au înrolat peste 50.000 de subiecți de caz.
   (c) De ce este nevoie de un număr atât de mare de subiecte?
6. (a) Ce factor este mai important în cauza bolilor: natura sau hrănirea? De ce?
   (b) Dintre acești doi factori, care este mai ușor de tratat/de gestionat? De ce?
7. (a) Cum pot fi luate în considerare diferențele dintre factorii genetici și factorii de mediu în investigațiile EPDM?
   (b) Dintre acești doi factori — genetici și de mediu — care este mai important? De ce?
8. Dacă o anumită abordare alternativă sau neortodoxă de gestionare a bolii (cum ar fi MTC, ayurveda, acupunctura etc.) pare a fi eficientă în gestionarea anumitor boli (de exemplu, diabet zaharat de tip 1, dureri de cap cronice), investigațiile EPDM pot fi utilizate pentru a corela abordarea „neortodoxă" a medicinei de bază, bazate pe dovezi? De ce sau de ce nu?

1.4 BIOS ȘI EPDM

Biostatistică - un cuvânt combinat derivat din ***biologie*** și ***statistică*** și, uneori, numit și *biometrie* sau *biometrie* - este aplicarea *statisticilor* la un subiect din biologie. BIOS include proiectarea de experimente biologice, în special în medicină și științele sănătății; colectarea și analiza datelor din acele experimente; și interpretarea rezultatelor.

Pentru a înțelege aplicarea BIOS-ului, să folosim exemplul unei comunități ipotetice din Statele Unite, în care un epidemiolog încearcă să cuantifice efectul unei anumite boli, cum ar fi gripa porcină (numele comun folosit pentru virusul H1N1, o nouă tulpină de gripă A, pentru a o deosebi de gripa sezonieră), pentru a studia distribuția bolii în diferite regiuni. Obiectivele sunt:

1. Pentru a determina amploarea populației afectate de boală
2. Pentru a stabili cauzele potențiale

Ar trebui mai întâi să se determine *prevalența* bolii, definită ca fracția de subiecți afectați de acea boală. În primul rând, se poate lua în considerare estimarea prevalenței populației în condiția asumată în mod obișnuit: eșantionarea în care se ia în considerare un eșantion randomizat de N subiecți, obținând X cazuri. De asemenea, pentru a înțelege caracterul aleatoriu al apariției bolii în populație, un biostatistician de cercetare începe adesea să analizeze datele în termenii unui model probabilistic. Un bun exemplu de model probabilistic elementar este *modelul binom,*

care se caracterizează prin încercări care se încheie într-unul din două moduri: fie succes, fie eșec . (*Bi* = doi; de aici numele „ *binomial* " care se referă la cele două posibile

rezultate ale succesului sau eșecului.)

Astfel, pentru a determina prevalența subiecților infectați cu boală, se poate lua un eșantion aleatoriu de

N = 500 de subiecți într-o anumită comunitate și obțineți
X = 4 subiecți cu rezultate pozitive la un test de anticorpi

Cu aceste rezultate, biostatisticianul are în mână următorul model:

1. Un model binom de distribuție a bolii
2. O medie eșantion de p = (X/N) = (4/500) = 1/125 = 0,008, din care se pot face predicții biostatistice folosind modelul de distribuție binomial bine-cunoscut.

(Se discută mai multe despre diferite modele probabilistice în capitolul 5 despre teoria probabilității și biostatistica inferențială.)

Astfel, pentru o populație mare (în aceeași condiție de distribuție) de n = 1.000.000, s-ar folosi același model (presupus a fi aplicabil în mod rezonabil) pentru a obține numărul așteptat de cazuri de boală. În acest exemplu, aceasta este dată de valoarea așteptată E[X] a modelului de distribuție binomială:

$$E[X] = np = (1.000.000)\,(0,008) = 8.000.$$

În mod clar, au fost făcute ipoteze majore. De exemplu, sa presupus că (a) modelul de distribuție binomială a fost aplicabil; și (b) a fost utilizată o singură populație de eșantionare.

În exemplul simplu de mai sus, abordarea epidemiologică a constat în:

În primul rând — Ipoteza unei distribuții de probabilitate pentru populație; în acest caz, distribuția binomială simplă

În continuare —Efectuarea unei eșantionări a populației și obținerea parametrilor de eșantionare pentru populație

În sfârșit — Utilizarea modelului probabilistic asumat pentru a face predicții cu privire la întreaga populație.

Aceștia sunt pașii clasici „1–2–3” în aplicarea **biostatisticii inferențiale** , pentru a trage concluzii prin abordarea inferenței. Desigur, pot fi utilizate modele biostatistice mai adecvate și poate fi necesară o eșantionare suplimentară, ceea ce duce la modele mai reprezentative și modele biostatistice mai rafinate. Este clar că concomitent cu dezvoltarea și utilizarea unor modele mai bune sunt mai implicate proceduri de calcul.

Pentru a sprijini astfel de calcule, software-ul open-source `R` va fi folosit în această călătorie a medicinei, medicinei preventive, sănătății publice, EPDM și BIOS.

În plus față de biostatistica inferențială, o abordare alternativă și mai simplă, numită *biostatistică descriptivă* , este adesea folosită în cercetarea epidemiologică. Obiectivul biostatisticii descriptive este pur și simplu de a descrie un set de date prin rezumarea tuturor caracteristicilor sale relevante. Ambele metode sunt descrise complet în Capitolul 2 .

Întrebări de revizuire pentru Secțiunea 1.4

1. În investigațiile EPDM, eșantionarea cantitativă este luată de la o populație țintă. Această abordare necesită metodologii preluate din BIOS? De ce sau de ce nu?
2. (a) Care este modelul de distribuție binomială în BIOS?
   (b) Cum ajută acest model la analiza datelor colectate într-o investigație EPDM?
3. (a) Este valoarea așteptată aceeași cu valoarea medie?
   (b) De ce sau de ce nu?
4. Care sunt pașii „1–2–3” în aplicarea biostatisticii inferențiale?
5. (a) Ce este biostatistica descriptivă?
   (b) Cum ajută biostatistica descriptivă în investigațiile EPDM?

REFERINȚE

Broadbent, A. (2009). Cauzele și modelele bolii în epidemiologie. *Studii în istoria și filosofia științelor biologice și biomedicale, 40,* 302–311. Preluat de pe http://www.hps.cam.ac.uk/people/broadbent/models_of_disease.pdf

Centre pentru Controlul și Prevenirea Bolilor. (2006). *Principiile epidemiologiei în practica sănătății publice: o introducere în epidemiologia aplicată și biostatistică* (ed. a treia). Atlanta, GA: Departamentul de Sănătate și Servicii Umane din SUA [Curs de auto-studiu SS1000].

Charlton, BG (2001). Libertatea personală sau sănătatea publică? În M. Marinker (Ed.), *Medicină și umanitate* (pp. 55–69). Londra: King's Fund. Preluat de la http://www.hedweb.com/bgcharlton/healthfreed.html

Edelman, D., Olsen, MK, Dudley, TK, Harris, AC și Oddone, EZ (2004). Utilitatea hemoglobinei A1c în prezicerea riscului de diabet. *Journal of General Internal Medicine (JGIM)* , *19* (12), 1175–1180.

Liu, X., Liu, L., Chen, P., Zhou, L., Zhang, Y., Wu, Y., . . . Yi, D. (2014). Studii clinice ale medicinei tradiționale chineze în tratamentul nefrologiei diabetului zaharat - O revizuire sistematică bazată pe o analiză de subgrup. *Jurnalul de etnofarmacologie* , *151* (2), 810–819.

Nathan, DM (2014). *Studiul pentru controlul și complicațiile diabetului/Epidemiologia intervențiilor și a complicațiilor diabetului la 30 de ani: Prezentare generală* . Preluat de pe http://care.diabetesjournals.org/

White, E., Armstrong, BK și Saracci, R. (2008). *Principii de măsurare a expunerii în epidemiologie : Colectarea, evaluarea și îmbunătățirea măsurilor factorilor de risc de boală* (ed. a 2-a). Oxford, Marea Britanie: Oxford University Press.

TWO

Cercetare și proiectare în epidemiologie și sănătate publică

INTRODUCERE

În cadrul unui program de sănătate publică (PH), cercetarea și proiectarea în epidemiologie (EPDM) ar trebui să se concentreze pe următoarele aspecte ale disciplinei (Universitatea Loma Linda, Scoala de Sanatate Publica, 2012) :

1. Efectuarea cercetării epidemiologice de calitate, inclusiv proiectarea adecvată, analiza biostatistică a datelor și interpretarea și raportarea rezultatelor
2. Realizarea si evaluarea studiilor clinice
3. Efectuarea supravegherii bolilor, așa cum se practică în departamentele de sănătate guvernamentale locale
4. Revizuirea critică a literaturii profesionale și identificarea punctelor forte și slabe ale proiectelor, analizelor și concluziilor
5. Evaluarea efectelor potențialelor confuzii și interacțiuni într-un proiect de cercetare
6. Aplicarea cunoștințelor despre mecanismele bolii și a informațiilor din disciplinele biologice la interpretarea constatărilor statistice în cercetarea biomedicală
7. Colaborarea cu profesioniștii din domeniul sănătății prin furnizarea de expertiză tehnică în ceea ce privește revizuirea literaturii de specialitate, proiectarea studiului, analiza datelor și interpretarea și raportarea rezultatelor
8. Menținerea unui mediu de calcul biostatistic rentabil care poate susține cererea și complexitatea tot mai mare de calcul științific

Această carte se concentrează pe limbajul de programare R , în sprijinul ultimului dintre aceste aspecte ale cercetării epidemiologice și promovează cele mai recente resurse pentru a încuraja un mediu de calcul construit în jurul utilizării lui R . Sunt prezentate mai multe exemple de lucru, folosind date de cercetare din viața reală obținute din investigații epidemiologice de ultimă oră și evidențiind programe și pachete R care sunt dezvoltate special pentru a procesa aceste seturi de date. Exemple de lucru suplimentare, selectate din alte aplicații biomedicale practice, sunt furnizate pentru a ilustra utilizarea mai largă a programării R în asociere cu alte limbaje de programare științifică de nivel înalt, cum ar fi FORTRAN.

Autorul intenționează ca cititorii să *repete mai întâi* calculele R prezentate aici. Aceasta este o modalitate de a obține o practică valoroasă și de a dezvolta încrederea într-un mediu avansat de cercetare bio -statistică .

2.1 CAUZAREA SI ASOCIAREA IN EPIDEMIOLOGIE SI SANATATE PUBLICA

EPDM urmărește să evalueze cauza bolii. Cu toate acestea, deoarece majoritatea studiilor epidemiologice sunt mai degrabă descriptive sau observaționale decât experimentale sau analitice , o serie de explicații posibile pentru o asociere observată trebuie luate în considerare înainte de a putea deduce în mod valid că există o relație cauză-efect. Adică, o asociere observată se poate datora de fapt efectelor unuia sau mai multor dintre următoarele (Rothman, 1998, 2002) :

- Șansă (eroare aleatorie)
- Prejudecăți (eroare sistematică)
- Derutant

Prin urmare, o asociere statistică observată între un factor de risc și o boală nu permite neapărat să se deducă o relație cauzală. În schimb, absența unei asocieri nu indică în mod concludent absența unei relații de cauzalitate. Judecata dacă o asociere statistică observată reprezintă o relație cauză-efect între expunere și boală necesită investigații și inferențe mult dincolo de datele dintr-un singur studiu; implică luarea în considerare a criteriilor care includ amploarea asocierii, consistența constatărilor din alte studii și credibilitatea biologică (Loma Linda University, School of Public Health, 2012 ; Steiger, 2015) .

În EPDM, criteriile Bradford-Hill sunt utilizate pe scară largă pentru a oferi un cadru în care să se evalueze dacă o asociere observată este probabil să fie cauzală.

Criteriile Bradford-Hill pentru cauzalitate și asociere în Epidemiologie (Hill, 1965)[6]

Setul de criterii de cauzalitate Bradford-Hill conturează condițiile minime necesare pentru a stabili o relație cauzală între doi itemi. Aceste criterii au fost prezentate inițial de Austin Bradford Hill (1897–1991) , un statistician medical britanic, ca o modalitate de a determina legătura cauzală dintre un anumit factor (de exemplu, fumatul de țigară) și o boală (cum ar fi emfizemul sau cancerul pulmonar) . Deși este ușor să pretindem că agentul „A" (de exemplu, fumatul) provoacă boala „B" (cancer pulmonar), este cu totul altă chestiune să stabilim o legătură semnificativă, biostatistic validă între cele două fenomene. Astfel, acest set de criterii a stat la baza cercetărilor epidemiologice moderne care stabilesc legături cauzale valide științific între potențialii agenți ai bolii și numeroasele boli care afectează omenirea. De fapt, aceste principii formează baza evaluării utilizate în *toate* cercetările științifice moderne.

Criteriile **Bradford-Hill** sunt prezentate aici, deoarece au fost aplicate în cercetarea epidemiologică:

1. ***Relație temporală.*** Expunerea precede întotdeauna rezultatul. Dacă se crede că factorul A cauzează o boală, atunci acel factor A trebuie să precedă întotdeauna apariția bolii. Acesta este singurul criteriu absolut esențial.
2. ***Puterea asocierii.*** Acest criteriu este definit de mărimea asociației, măsurată prin teste biostatistice adecvate. Cu cât asocierea este mai puternică, cu atât este mai probabil ca relația dintre A și B să fie cauzală. De exemplu, cu cât hipertensiunea

[6] www.drabruzzi.com/hills_criteria_of_causation.htm

arterială (tensiune arterială crescută) este mai puternic corelată cu o dietă bogată în sodiu, cu atât este mai puternică relația dintre hipertensiune arterială și aportul de sodiu.

3. ***Relația doză-răspuns*** . Pe măsură ce cantitatea de expunere crește, crește și riscul. Dacă este prezentă o relație doză-răspuns, aceasta este o dovadă puternică a unei relații cauzale. Cu toate acestea, ca și în cazul specificității (a se vedea punctul 8 din această listă), absența unei relații doză-răspuns nu exclude neapărat o relație cauzală; poate exista un prag peste care se dezvoltă anumite relaţii specifice. În același timp, dacă un anumit factor este cauza unei boli, incidența bolii ar trebui să scadă atunci când expunerea la factor este redusă sau eliminată. De exemplu , în epidemiologia mediului, dacă nivelurile crescânde de dioxid de carbon din atmosferă determină o creștere a temperaturilor globale, atunci — cu celelalte lucruri fiind egale — ar trebui să ne așteptăm atât la o creștere proporțională, cât și la o scădere corespunzătoare a temperaturii globale în urma unei creșteri sau scăderi. , respectiv, în nivelurile de dioxid de carbon din atmosferă.
4. ***Consecvență*** . O asociere este *consistentă* atunci când rezultatele sunt replicate în studii în diferite populații folosind metode diferite. Astfel, dacă o relație este cauzală, ne-am aștepta să apară în mod constant în diferite studii și în rândul diferitelor - populații. Prin urmare, trebuie întreprinse numeroase experimente înainte de a putea fi făcute declarații semnificative despre relația cauzală dintre doi sau mai mulți factori. De exemplu, a fost nevoie de multe studii științifice riguroase ale relației dintre fumatul de țigară și cancer înainte de a se putea ajunge la concluzia definitivă că fumatul de țigară crește riscul de cancer – dar rețineți că încă nu putem afirma că „fumatul provoacă cancer”.
5. ***Plauzibilitatea*** . Acest criteriu necesită ca asociația să fie de acord cu înțelegerea acceptată în prezent a proceselor patologice. Adică, trebuie să existe o bază teoretică pentru a afirma o asociere între un comportament de stil de viață și o anumită boală. În același timp, cercetările care nu sunt de acord cu teoria stabilită nu sunt neapărat false sau eronate. Mai degrabă, poate indica faptul că este justificată o reconsiderare a principiilor acceptate în prezent.
6. ***Luarea în considerare a explicațiilor alternative*** . Pentru a decide dacă o asociere raportată este cauzală, este necesar să se determine măsura în care anchetatorii au luat în considerare alte explicații posibile și au exclus efectiv acele explicații alternative. Cu alte cuvinte, este întotdeauna necesar să se ia în considerare mai multe ipoteze înainte de a trage concluzii despre relația cauzală dintre oricare doi elemente investigate.
7. ***Experimentează*** . Condiția poate fi schimbată, slăbită sau prevenită printr-o abordare experimentală adecvată.
8. ***Oraș specific*** . Aceasta se stabilește atunci când o singură cauză general acceptată produce un efect specific. Acesta este considerat de unii ca fiind cel mai slab dintre toate criteriile. De exemplu, bolile atribuite fumatului nu îndeplinesc acest criteriu. Când se găsește specificitatea unei asocieri, aceasta oferă un suport suplimentar pentru o relație cauzală. Cu toate acestea, absența specificității nu anulează în niciun caz o relație cauzală, deoarece anumite rezultate, cum ar fi răspândirea unei boli, sunt susceptibile de a avea mai mulți factori cauzali și influențe și este foarte important.
   este puțin probabil ca cineva să găsească o relație unu-la-unu cauză-efect între două fenomene. Cauzalitatea este cel mai adesea multiplă; prin urmare, este necesar să se examineze relațiile cauzale specifice într-o perspectivă mai largă.
9. ***Coerență*** . Asocierea ar trebui să fie compatibilă cu teoriile și cunoștințele existente.

Adică, este necesară evaluarea pretențiilor de cauzalitate în contextul stării actuale a cunoștințelor, atât în cadrul unui domeniu dat, cât și în domenii conexe. Altfel spus, ce trebuie să sacrifice despre ceea ce știm în prezent pentru a accepta o anumită pretenție de cauzalitate? De exemplu, ce teorii și principii acceptate în prezent în antropologie, biologie, chimie și fizică trebuie să respingem pentru a accepta afirmația creaționistă că lumea a fost creată exact și literal așa cum este descris în Biblie? Cu toate acestea, ca și în problema plauzibilității, *rezultatele cercetării care sunt în dezacord sau în conflict cu teoria și cunoștințele stabilite nu sunt automat greșite sau false* . Ei pot forța o reconsiderare a credințelor și principiilor acceptate. Toți cei acceptați în prezent , inclusiv teoria evoluționistă, teoria relativității și teoria ecologiei populației, au fost la un moment dat idei noi care au contestat ortodoxia. Astfel de schimbări în teoriile acceptate sunt cunoscute și sub numele de *schimbări de paradigmă* .

COMENTARIILE LUI ROTHMAN LA CRITERIILE BRADFIELD-HILL (ROTHMAN, 1998, 2002)

Criticii au susținut că criteriile Bradford-Hill nu reușesc să ofere speranța de a distinge clar relațiile cauzale de cele noncauzale. De exemplu, al doilea criteriu , puterea asocierii, nu ține cont de faptul că nu toate cauzele componente vor avea o asociere puternică cu boala pe care o produce și nici nu recunoaște că puterea asocierii depinde adesea de prevalența altora. factori.

Al treilea criteriu, relația doză-răspuns, sugerează că o asociere cauzală este mai probabilă dacă se poate demonstra un gradient biologic sau o curbă doză-răspuns . Cu toate acestea, astfel de relații pot rezulta din confuzie sau alte părtiniri. Potrivit lui Rothman, singurul criteriu care este cu adevărat un criteriu cauzal este temporalitatea – adică o determinare că cauza a precedat efectul. (Totuși, poate fi dificil să se stabilească succesiunea de timp a cauzei și efectului.)

Al optulea criteriu, specificitatea, care sugerează că o relație este mai probabil să fie cauzală dacă expunerea este legată de un singur rezultat, poate induce în eroare, deoarece o cauză poate avea multe efecte. De exemplu, fumatul este asociat cu multe efecte neplăcute, de la degete pătate și respirație urât mirositoare până la cancer pulmonar.

În general, procesul de inferență cauzală este complex și ajungerea la o identificare tentativă a naturii cauzale sau noncauzale a unei asocieri este un proces surprinzător de subiectiv.

Interpretare juridică folosind epidemiologie[7]

EPDM poate stabili doar că un agent *ar fi putut* produce un efect în orice caz particular; nu poate dovedi că agentul a cauzat efectiv efectul. Această axiomă importantă stă la baza tuturor cercetărilor epidemiologice și biostatistice.

EPDM se preocupă de *incidența* bolilor în *populații* ; nu abordează întrebarea ce a cauzat boala unui anumit individ. Această întrebare , denumită uneori *cauzalitate specifică* , este dincolo de domeniul științei epidemiologiei. EPDM are limite în punctul în care se face o inferență sau o afirmație că relația dintre un agent și o boală este cauzală și în care a fost determinată amploarea riscului excesiv atribuit agentului. Adică, EPDM

[7] http://en.wikipedia.org/wiki/Epidemiology

abordează dacă un agent *poate* provoca o boală, nu dacă un agent *a* provocat o anumită instanță a unei boli.

Cu toate acestea, subdisciplina **epidemiologiei criminalistice** este îndreptată spre investigarea cauzei specifice a bolii sau rănilor la un individ sau un grup de indivizi în cazurile în care cauzalitatea este contestată sau neclară, pentru prezentare în cadrul juridic. Conform legislației SUA, EPDM singur nu poate dovedi că o asociere cauzală *nu* există în general. În schimb, o instanță din SUA într-un caz individual specific poate accepta dovezi epidemiologice pentru a susține sau a justifica o deducere că există o asociere cauzală, bazată pe echilibrul probabilității.

Apariția bolilor (CDC, 2006)

O poziție critică luată în EPDM este că bolile nu apar aleatoriu într-o populație , dar sunt mai probabil să apară la unii membri ai populației decât la alții din cauza *factorilor de risc* care nu sunt distribuiți aleatoriu în populație. EPDM este utilizat pentru a identifica factorii care pun unii membri ai unei populații la un risc mai mare decât alții.

CAUZALA

Au fost propuse o serie de modele de cauzalitate a bolii. Printre cele mai simple este triunghiul epidemiologic, modelul tradițional al bolilor infecțioase. Triunghiul este format din:

- **agent** extern
- O **gazdă** susceptibilă
- Un **mediu** care aduce împreună gazda și agentul

În acest model, boala rezultă din interacțiunea dintre agent și gazdă într-un mediu care susține transmiterea agentului de la o sursă la acea gazdă.

Agentul, gazda și factorii de mediu se pot interconecta în multe moduri complexe pentru a produce boli. Diferitele boli necesită echilibre și interacțiuni diferite ale acestor trei componente. Dezvoltarea unor măsuri adecvate, practice și eficiente de PH pentru controlul sau prevenirea bolii necesită de obicei evaluarea tuturor celor trei componente și a interacțiunilor lor.

Agentul poate fi un microorganism infecțios sau **un** agent patogen: un virus, o bacterie , un parazit sau alți microbi. Agentul trebuie să fie prezent pentru ca boala să apară; totuși, prezența acelui agent în sine nu este întotdeauna suficientă pentru a provoca boala. O varietate de factori influențează dacă expunerea la un organism va duce la boală , inclusiv capacitatea organismului de a provoca boală (cunoscută sub numele de *patogenitatea sa*) și nivelul dozei.

Conceptul de agent a fost extins pentru a include cauzele chimice și fizice ale bolii sau rănilor, cum ar fi contaminanții chimici. De exemplu, contaminarea suplimentelor nutritive de proastă calitate care conțin L-triptofan a fost considerată responsabilă pentru apariția sindromului eozinofilie-mialgie, o afecțiune neurologică incurabilă și uneori fatală asemănătoare gripei. (**Triptofanul** este un aminoacid esențial în dieta umană. Numai stereoizomerul L al triptofanului este utilizat în proteinele structurale sau enzimatice, dar stereoizomerul D se găsește ocazional în peptidele produse în mod natural.)

Deși triunghiul epidemiologic (sau triada) servește ca un model util pentru multe boli, s-a dovedit inadecvat pentru boli cardiovasculare, cancer și alte boli care par să aibă mai multe cauze care contribuie fără una singură necesară.

O varietate de factori intrinseci gazdei individuale (omul care poate face boala), numiți uneori *factori de risc* , pot influența expunerea individului, susceptibilitatea și răspunsul la un agent cauzal. Oportunitățile de expunere sunt adesea influențate de comportamente precum igiena, practicile sexuale și alte opțiuni personale, precum și de vârstă și sex. Susceptibilitatea și răspunsul la un agent sunt influențate de factori precum genetica, starea nutrițională și imunologică, structura anatomică, prezența altor boli sau medicamente și structura psihologică.

Mediul se referă la factori extrinseci care afectează atât agentul, cât și oportunitatea de expunere. Factorii de mediu includ factori fizici, cum ar fi geologia și clima, factorii biologici, cum ar fi insectele care transmit agentul și factorii socioeconomici, cum ar fi aglomerația, salubritatea și disponibilitatea serviciilor de sănătate.

CAUZELE COMPONENTE ŞI PÂINTE CAUZALE

Deoarece modelul agent-gazdă-mediu nu funcționează bine pentru multe boli neinfecțioase, au fost propuse câteva alte modele care încearcă să țină seama de natura multifactorială a cauzalității. Un astfel de model, propus de Rothman în 1976, a ajuns să fie cunoscut sub numele de „plăcinte cauzale" (Rothman, 1976).

Acest model este ilustrat în Figura 2.1 . Un factor individual care contribuie la provocarea bolii este prezentat ca o bucată de plăcintă. După ce toate bucățile unei plăcinte sunt asamblate , plăcinta este completă și apare boala. Factorii individuali (bucăți din plăcintă) se numesc **cauze componente** . Placinta completă, care ar putea fi considerată o cale cauzală, se numește o **cauză suficientă** . O boală poate avea mai mult de o cauză suficientă, fiecare cauză suficientă fiind compusă din mai multe cauze componente care se pot suprapune sau nu. O componentă care apare în fiecare fel de plăcintă sau cale se numește o **cauză necesară** , deoarece fără ea, boala nu apare.

Cauzele componente pot include factori gazdă intrinseci, precum și agentul și factorii de mediu din triunghiul agent-gazdă-mediu. O cauză cu o singură componentă este rareori o cauză suficientă în sine. Susceptibilitatea gazdei și alți factori ai gazdei pot juca, de asemenea, un rol. De exemplu, chiar și expunerea la un agent extrem de infecțios, cum ar fi virusul rujeolic, nu duce invariabil la boala rujeolă. În schimb , un agent care este de obicei inofensiv la persoanele sănătoase poate provoca boli devastatoare în diferite condiții.

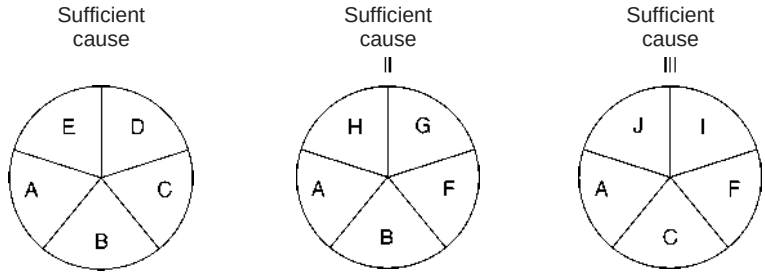


FIGURE 2.1 Rothman's causal pies.

De exemplu, *Pneumocystis carinii* (*P. carinii*) este un organism care poate coloniza tractul respirator al unor persoane sănătoase fără a provoca niciun rău, dar poate provoca pneumonie potențial letală la persoanele al căror sistem imunitar a fost compromis de HIV. Prezența organismelor *P. carinii* este, prin urmare, o cauză *necesară, dar nu suficientă,* a pneumoniei pneumocystis. În Figura 2.1 , această situație este reprezentată de cauza componentei A. După cum indică modelul, o anumită boală poate rezulta dintr-o varietate de cauze sau căi suficiente diferite.

Ca un alt exemplu, cancerul pulmonar poate rezulta dintr-o cauză suficientă care include fumatul ca cauză componentă. Cu toate acestea, deoarece nu toți fumătorii dezvoltă cancer pulmonar, fumatul în sine nu poate fi identificat ca o cauză suficientă. Mai mult, deoarece un număr mic de pacienți cu cancer pulmonar nu au fumat niciodată, fumatul nu este nici o cauză necesară.

Să presupunem că componenta cauza B este fumatul și componenta cauza C este expunerea la azbest:

- Cauza suficientă I include atât fumatul (B), cât și expunerea la azbest (C).
- Cauza II suficientă include expunerea la azbest, dar nu fumatul.
- Cauza III suficientă include fumatul fără expunere la azbest.

Deoarece cancerul pulmonar se poate dezvolta la persoane care nu au fost niciodată expuse nici la fum sau la azbest, un model adecvat pentru cancerul pulmonar ar trebui să arate cel puțin o plăcintă cu cauze suficiente care nu include nici componenta B, nici componenta C.

Totuși, acțiunea PH nu depinde de identificarea fiecărei cauze componente. Prevenirea bolilor poate fi realizată prin blocarea *oricărei* singure componente a unei cauze suficiente, cel puțin prin acea cale. De exemplu, eliminarea fumatului (componenta B) ar preveni cancerul pulmonar care decurge din cauze suficiente I și II, deși unele cazuri de cancer pulmonar ar apărea totuși prin cauza III suficientă.

FORTA UNEI CAUZE

În EPDM, **puterea** unui factor poate fi măsurată prin modificarea frecvenței bolii - produsă prin introducerea factorului într-o populație. Această modificare poate fi măsurată în termeni absoluti sau relativi. În ambele cazuri, puterea unui efect poate avea o semnificație importantă concomitentă a PH, dar o semnificație biologică mică; aceasta deoarece, având în vedere un mecanism cauzal specific, oricare dintre cauzele componente poate avea efecte puternice sau slabe. Identitatea reală a componentelor constitutive ale mecanismului cauzal echivalează cu biologia cauzalității.

În plus, puterea efectului unui factor depinde de distribuția specifică în timp a

complementelor sale cauzale în populație. De-a lungul unei perioade de timp, puterea efectului unui anumit factor în ceea ce privește apariția unei anumite boli se poate schimba, deoarece prevalența complementelor sale cauzale în diferite mecanisme cauzale se poate modifica, de asemenea. Acest lucru poate apărea chiar și atunci când mecanismele cauzale prin care acționează factorul și complementele sale rămân neschimbate.

INTERACȚIUNE ÎNTRE CAUZE

Modelul plăcintă cauzală recunoaște că mai multe componente cauzale pot acționa în comun pentru a produce un efect. Cu toate acestea, „acționarea concertată" nu implică neapărat că toți factorii trebuie să acționeze în același timp.

Luați în considerare exemplul unei persoane care suferă o rănire la cap care are ca rezultat o tulburare de echilibru; ani mai târziu, această persoană cade pe o potecă alunecoasă și își rupe șoldul. Leziunile capului anterioare și problema de echilibru au jucat un rol *cauzal* în fractura de șold rezultată, dar la fel și condițiile meteorologice din ziua fracturii. Dacă ambii acești factori au jucat un rol cauzal în fractura de șold, atunci ei au interacționat unul cu celălalt pentru a provoca fractura, în ciuda faptului că perioadele lor de acțiune au fost la mulți ani. Acest lucru permite concluzia că *oricare dintre factorii din același mecanism cauzal al bolii interacționează unul cu altul pentru a provoca boala.*

Se poate vedea fiecare plăcintă cauzală ca un set de componente cauzale care interacționează. În exemplul precedent, leziunea anterioară a capului a interacționat cu condițiile meteorologice, precum și cu alte cauze componente, cum ar fi tipul de pantofi purtati, absența unei prinderi de mână și orice alte condiții care au fost necesare mecanismului cauzal al căderea și șoldul rupt care a rezultat. Astfel, acest model oferă o bază biologică pentru un concept de interacțiune între cauze.

Întrebări de revizuire pentru Secțiunea 2.1

1. Sugerați aspecte ale disciplinelor din cadrul EPDM care ar trebui să fie punctul central pentru un program de PH.
2. Când un investigator evaluează cauzele bolilor, ce efecte ar putea duce la o asociere greșită observată pe care investigatorul o consideră că indică existența unei relații cauză-efect?
3. (a) Care sunt criteriile Bradford-Hill pentru cauzalitate și asociere în EPDM?
   (b) În ce mod sunt realiste obiecțiile Rothman la criteriile Bradford-Hill în ceea ce privește următoarele?
   - Puterea asocierii
   - Relația doză-răspuns
   - Specificitate
4. Opinia juridică este că „epidemiologia poate stabili doar că un agent ar fi putut cauza, dar nu că a provocat, un efect într-un anumit caz". Este aceasta o - interpretare rezonabilă? De ce sau de ce nu?
5. (a) Ce este triunghiul epidemiologic?
   (b) Ilustrați schematic acest concept folosind două modele diferite.

6. În contextul cauzei bolilor, ce este
   (a) o cauza necesara?
   (b) o cauza suficienta?
   Dați un exemplu pentru fiecare.
7. Discutați conceptul de plăcinte cauzale lui Rothman și dați câteva exemple.
8. În contextul cauzei bolii, discutați noțiunea de cauză „necesară, dar nu suficientă" a unei boli. Dați câteva exemple.

2.2 CAUZAREA ŞI INFERENŢA ÎN EPIDEMIOLOGIE ȘI SĂNĂTATE PUBLICA

În EPDM și PH (Rothman & Greenland, 2005), o **cauză** a unei boli este definită ca un eveniment, afecțiune sau caracteristică care *a precedat* boala, fără de care boala fie nu ar fi apărut, fie nu ar fi apărut decât mai târziu. . În mod remarcabil, conform acestei definiții, este posibil ca niciun eveniment, condiție sau caracteristică specifică să nu fie suficientă în sine pentru a produce (cauza) boala.

O *cauză suficientă* este un mecanism cauzal complet care poate fi definit ca un set de condiții și evenimente minime care produc inevitabil boală. Aici, *minim* implică faptul că toate condițiile sau evenimentele sunt necesare pentru acea apariție. În studiul cauzei bolii, completarea sau existența unei cauze suficiente poate fi considerată echivalentă cu debutul bolii. (*Debutul* se referă la stadiul cel mai timpuriu al procesului bolii, mai degrabă decât la apariția reală a simptomelor.)

Uneori, toate componentele unei cauze suficiente pentru un anumit efect biologic sunt necunoscute. De exemplu, fumatul este o cauză a cancerului pulmonar, dar în sine nu este o cauză suficientă. În primul rând, termenul *de fumat* este prea imprecis pentru a fi folosit într-o descriere cauzală. Astfel, trebuie precizat:

- Tipul de fum (de exemplu, țigară, trabuc, pipă)
- Indiferent dacă fumul este filtrat sau nefiltrat
- Modul și frecvența inhalării
- Debutul și durata fumatului

Acești *factori de susceptibilitate* se referă și interacționează cu alte componente în diferitele mecanisme cauzale prin care fumatul provoacă cancer pulmonar.

De asemenea, rețineți că fumatul nu va provoca cancer la toți cei care fumează. Aparent, există unii oameni care, în virtutea componenței lor genetice sau a experienței anterioare, sunt susceptibili la efectele nocive ale fumatului, iar alții care nu sunt (uneori sunt cunoscuți ca *șobolanii duri,* cu referire la experimentele pe animale de laborator) .

Diagramele lui Rothman pentru o cauzare suficientă a Boli (Broadbent, 2011)

O diagramă schematică a cauzelor suficiente într-o boală este prezentată în Figura 2.2 (Rothman & Groenlanda, 2005) . În această diagramă:

- Fiecare set de cauze componente reprezentate este minim suficient pentru a produce boala.

- Fiecare cauză componentă este o parte necesară a acelui mecanism cauzal specific .
- O anumită cauză componentă poate juca un rol în unul, două sau toate cele trei mecanisme cauzale prezentate.

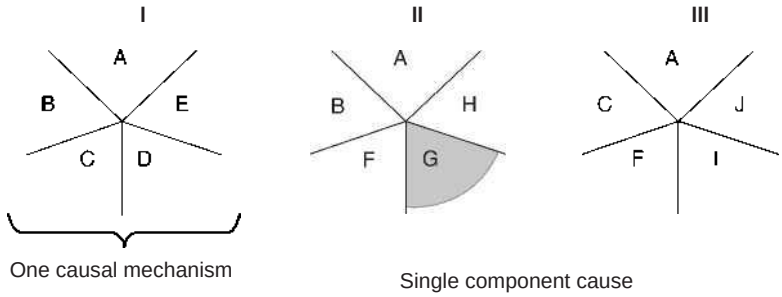


FIGURA 2.2 Diagramele lui Rothman pentru cauze suficiente de boli.

Diagrama de cauzalitate prezentată în Figura 2.2 evidențiază câteva principii importante în ceea ce privește **inferența** și **cauzalitatea** în bolile umane. Primul, și poate cel mai important, dintre aceste principii este evident din diagramă: O anumită boală poate avea mai multe cauze și fiecare mecanism cauzal implică acțiunea comună a unei multitudini de cauze componente.

■ **Exemplu:** Acest lucru aduce înapoi la exemplul nostru anterior al cauzei unui șold rupt. John suferă o accidentare gravă la cap care duce la o perturbare permanentă a echilibrului . Ani mai târziu, acel echilibru defectuos joacă un rol cauzal într-o cădere care are loc în timp ce John merge pe o potecă alunecoasă. Acea cădere are ca rezultat ruperea șoldului lui John. Alți factori care joacă un rol cauzal în șoldul rupt ar putea include:

1. Tipul de pantof pe care îl purta John
2. Lipsa balustradei de-a lungul potecii
3. O rafală de vânt neașteptat de puternică
4. Greutate corporală excesivă (John este oarecum obez)

Toate acestea, pe lângă mulți alți factori, ar putea juca roluri cauzale.

Astfel, mecanismul cauzal complet implică mulți factori. Unii factori, cum ar fi greutatea lui John și rănirea anterioară care a dus la perturbarea echilibrului său, reprezintă evenimente sau condiții anterioare care au avut efecte persistente. Unele componente cauzale genetice ar fi putut afecta greutatea lui John, felul lui de a merge, comportamentul, recuperarea lui de la accidentarea capului anterioară și așa mai departe. Alți factori, cum ar fi forța rafalei bruște de vânt, sunt de mediu. Putem afirma în mod rezonabil că există unele cauze genetice și unele componente de mediu în fiecare mecanism cauzal. Astfel, chiar și un eveniment, cum ar fi o cădere pe o potecă alunecoasă, care are ca rezultat o ruptură de șold, face parte dintr-un mecanism cauzal complicat care implică multe cauze componente.

Punctul important despre multicauzalitate este că majoritatea cauzelor identificate nu sunt nici necesare, nici suficiente pentru a produce boala. Cu toate acestea, o cauză nu trebuie să fie necesară sau suficientă pentru ca eliminarea ei să aibă ca rezultat prevenirea bolii. Lucrătorii PH știu că dacă o cauză componente care nu este nici necesară, nici suficientă este eliminată,

o cantitate substanțială de boală poate fi încă prevenită. Faptul că o cauză nu este necesară - înseamnă că o boală poate apărea în continuare după eliminarea acestei cauze, dar acea cauză

componente va fi totuși o cauză necesară pentru *unele* dintre cazurile care apar. Faptul că cauza componentei nu este suficientă implică faptul că (a) alte cauze componente trebuie să interacționeze cu aceasta pentru a produce boala și (b) blocarea oricăreia dintre ele va duce la prevenirea unor cazuri de boală. Astfel, nu este nevoie să identificăm fiecare cauză componente pentru a preveni unele cazuri de boală. Eforturile de PH care abordează doar câteva dintre cauzele componente aparent cele mai importante pot avea succes semnificativ în reducerea incidenței unei boli; face parte din sarcina EPDM de a ajuta la identificarea acestor cauze importante.

ÎNSUMAREA FRACȚIUNILOR ATRIBUIBILE ALE CAUZELOR DE BOLI

Astfel, se poate vedea fiecare plăcintă cauzală ca un set de componente cauzale care interacționează. Acest model oferă o bază biologică pentru un concept de interacțiune care este diferit de viziunea biostatistică obișnuită a interacțiunii (Ephron, 1984). Luați în considerare setul de date ipotetic din Tabelul 2.1 .

În situația de fumat și cancer de cap/gât, diferențele dintre rate reflectă toate efectele cauzale. Dintre acele persoane care sunt fumători de țigări și, de asemenea, băutori de alcool, ce proporție din cazuri este atribuită efectului fumatului?

În acest exemplu, se dă că rata pentru aceste persoane este de 15 cazuri la 100.000 persoane-ani. Dacă acești oameni nu ar fi fumători de țigări, putem deduce că rata lor de cancer la cap și gât ar fi de 4 cazuri la 100.000 de persoane-an. Dacă această diferență reflectă *rolul cauzal al fumatului de țigară* , atunci am putea deduce că (15 - 4) = 11 din 15 cazuri, sau 73,3%, sunt atribuite fumatului de țigări în rândul celor care *fumează* și beau alcool.

Dacă întrebarea este ce proporție de boală în rândul acelorași oameni este atribuită - consumului de alcool, am putea atribui (15 – 5) = 10 din 15 cazuri, sau 66,7%, *rolului cauzal al consumului de alcool* .

Acum, se poate atribui 73,3% din cazuri fumatului și 66,7% consumului de alcool în rândul celor care sunt expuși la ambele? Da, pentru că unele cazuri sunt numărate de mai multe ori.

De asemenea, fumatul și alcoolul interacționează în unele cazuri de cancer la cap/gât, iar aceste cazuri sunt atribuite atât fumatului de țigări, cât și consumului de alcool. O consecință a interacțiunii este că proporțiile bolii atribuibile diferitelor cauze ale componentelor se însumează rareori exact la 100%.

TABELUL 2.1 Ratele ipotetice ale cancerului de cap/gât în funcție de starea de fumat și de consumul de alcool

BĂUT ALCOOL	NU	DA
STARE DE FUMĂT DE TIGARĂ		
Nefumător	1	4
Fumător	5	15

Notă: Cazuri la 100.000 persoane-ani

În anii 1970 (Ephron, 1984) , oamenii de știință de la Institutul Național de Sănătate din SUA (NIH) au propus că până la 40% din cancer este atribuită expunerilor profesionale . Mulți au considerat că această fracțiune este o supraestimare. Unul dintre argumentele folosite în contracție a fost următorul: *x* % din cancer este cauzat de fumat, *y* % de dietă, *z* % de alcool și așa mai departe. Când se adună toate aceste procente, doar un mic procent (mult mai puțin de 40%) rămâne pentru cauze profesionale. Cu toate acestea, această respingere este greșită, deoarece se bazează pe premisa incorectă că

fiecare caz de boală are o singură cauză și că două sau mai multe cauze nu pot contribui ambele la același caz de cancer. Astăzi, se acceptă faptul că, deoarece stilul de viață (dieta, exercițiile fizice, fumatul, consumul de alcool etc.), mediul (poluarea aerului, expunerea la azbest etc.) și diferitele expuneri profesionale, împreună cu alți factori, interacționează între ele și cu factori genetici (ereditari) care cauzează cancer, fiecare caz de cancer poate fi atribuit *în mod repetat* mai multor cauze componente separate.

Suma bolilor atribuibile diverselor cauze componente nu are astfel o limită superioară. O singură cauză sau categorie de cauze care este prezentă în fiecare cauză suficientă de boală va avea o fracțiune atribuibilă de 100%.

Au fost întreprinse multe cercetări pentru a dezvolta „indici de ereditare", care ar trebui să măsoare fracția de boală care este moștenită. Cu toate acestea, acești indici evaluează doar rolul relativ al cauzelor de mediu și genetice ale bolii într-un anumit cadru. De exemplu, unele cauze genetice pot fi componente necesare oricărui mecanism cauzal. Dacă toată lumea dintr-o populație are un set identic de gene care cauzează boala, atunci efectul lor nu este inclus sau reflectat în indicii de ereditare, chiar dacă acest set de gene este o cauză a bolii. Dacă sunt luați în considerare toți factorii genetici care determină boala, indiferent dacă aceștia variază sau nu în cadrul populațiilor, atunci se poate spune că 100% din boală este moștenită. În mod analog, 100% din orice boală este cauzată de mediu, chiar și acele boli pe care adesea le considerăm pur genetice.

■ **Exemplul 1:** Fenilcetonuria este considerată de mulți ca fiind pur genetică. Cu toate acestea, retardul mintal pe care îl poate provoca poate fi prevenit printr-o intervenție alimentară adecvată. Tratamentul pentru fenilcetonurie ilustrează interacțiunea dintre gene și mediu pentru a provoca o boală considerată în mod obișnuit a fi pur genetică.

■ **Exemplul 2:** Cum rămâne cu o cauză a decesului aparent pur de mediu , cum ar fi decesul dintr-un accident de mașină? Este ușor de conceput trăsături genetice care duc la probleme psihice, cum ar fi alcoolismul, care, la rândul lor, duc la conducerea în stare de ebrietate și, în consecință, accidente de mașină fatale. În acest exemplu, interacțiunea dintre gene și mediu provoacă o boală considerată în mod obișnuit a fi pur de mediu.

■ **Exemplul 3:** Luați în considerare exemplul extrem de mediu de a fi ucis de fulger . Condițiile psihiatrice parțial ereditare pot influența dacă cineva se va adăposti în timpul unei furtuni cu fulgere; trăsăturile genetice, cum ar fi capacitatea atletică, pot influența probabilitatea de a fi în aer liber atunci când lovește o furtună cu fulgere; și a avea o ocupație sau distracție în aer liber care este mai frecventă în rândul bărbaților (și în acest sens este genetică) ar influența, de asemenea, probabilitatea de a fi ucis de un fulger. Mai mult, chimia corpului unor persoane – un factor genetic – îi poate face de fapt mai susceptibili la lovituri de fulger.

Aceste exemple arată clar că fiecare caz de boală are cauze atât genetice, cât și de mediu. Acest mod de gândire este defensabil și are implicații importante pentru cercetare.

Inferențe cauzale

Inferența cauzală este un caz special al procesului mai general de raționament științific, despre care există o dezbatere substanțială între oamenii de știință și filozofi.

IMPOSIBILITATEA DOVENDII

Filosoful David Hume a observat în secolul al XVIII-lea că dovezile sunt imposibile în știința empirică, iar observația sa este încă adevărată. Epidemiologii se confruntă adesea cu criticile conform cărora dovada este imposibilă în EPDM - cu implicația că este posibilă în alte discipline științifice! Unii oameni de știință experimental susțin că relațiile epidemiologice pot fi doar sugestive și că studiul detaliat de laborator al mecanismelor la câțiva indivizi poate dezvălui relațiile cauză-efect cu certitudine.

Din nefericire, această viziune trece cu vederea faptul că toate relațiile sunt sugestive exact în modul descris de Hume: Chiar și cea mai atentă și detaliată analiză a evenimentelor individuale nu poate oferi mai mult decât asocieri, deși la niveluri foarte detaliate . Studiile de laborator implică un grad de control al observatorului care nu poate fi egalat în EPDM; doar acest control, nu nivelul de observație, se presupune că stă la baza derivării „dovadă" din studiile de laborator.

Cu toate acestea, un astfel de control nu este o garanție împotriva erorilor. Toată munca științifică, în EPDM sau în orice altă disciplină, este în cel mai bun caz doar o descriere tentativă a naturii și a realității.

TESTAREA TEORIILOR EPIDEMIOLOGICE CONCURENȚE

Cunoștințele biologice despre ipotezele epidemiologice sunt adesea limitate, iar acest fapt duce ocazional la ipoteze care par a fi puțin mai mult decât afirmații vagi de asociere cauzală între expunere și boală (de exemplu, „fumatul provoacă boli cardiovasculare"). Deși aceste ipoteze vagi par adesea intuitiv corecte, ele pot fi dificil de verificat empiric. Pentru a face față acestei provocări, epidemiologii se concentrează adesea pe testarea *negației* ipotezei cauzale; în termeni probabilistici, aceasta este echivalentă cu **ipoteza nulă** (H_0) că expunerea *nu* are o relație cauzală cu boala. În astfel de teste, orice asociere observată poate *respinge* ipoteza, presupunând că părtinirile sunt absente.

Dacă mecanismul cauzal este precizat suficient de specific, observațiile epidemiologice pot sugera teste cruciale ale ipotezelor cauzale concurente non-nule. Cu toate acestea, multe studii epidemiologice nu sunt concepute pentru a testa o ipoteză cauzală.

- **Exemplu:** Horwitz și Feinstein (1978) au examinat date epidemiologice timpurii care păreau să indice că femeile care au urmat terapie de substituție cu estrogeni au un risc mai mare de cancer endometrial. Acești cercetători au emis apoi ipoteza unei asocieri conform căreia femeile care iau estrogen vor prezenta simptome precum sângerare, care le-ar determina apoi să solicite asistență medicală. Efortul ulterior a avut ca rezultat un important

procedura de diagnosticare care a permis detectarea cancerului endometrial într-un stadiu mai devreme la aceste femei, în comparație cu femeile care nu iau estrogeni în primul rând.

Observațiile epidemiologice au fost utilizate pentru a evalua ipotezele concurente probabile:

- cauzală a prezis că riscul de cancer endometrial ar tinde să crească odată cu utilizarea crescută a estrogenilor.
- Teoria părtinirii de detectare a prezis că femeile care au folosit estrogeni doar pentru o perioadă scurtă de timp ar avea cel mai mare risc, deoarece simptomele legate de utilizarea estrogenului, care au condus la consultarea medicală precoce, au avut tendința să apară la scurt timp după începerea utilizării.

Deoarece asocierea dintre consumul recent de estrogen și cancerul endometrial a fost aceeași atât la utilizatorii de estrogen pe termen lung, cât și la cei pe termen scurt, teoria părtinirii de detecție a fost respinsă ca o explicație pentru toate, cu excepția unei mici fracțiuni, cazurile de cancer endometrial apărute după utilizarea estrogenului.

Acest exemplu ilustrează un punct critic în înțelegerea procesului de **inferență cauzală** în EPDM. Multe dintre ipotezele evaluate prin interpretarea cercetărilor epidemiologice sunt ipoteze noncauzale; adică nu implică nicio legătură cauzală între expunerea la studiu și boală. Astfel, ipotezele care explică modul în care anumite tipuri de părtinire ar fi putut duce la o asociere între expunere și boală sunt alternativele obișnuite la ipoteza studiului primar pe care epidemiologul trebuie să o ia în considerare pentru a ajunge la inferențe. O mare parte din interpretarea studiilor epidemiologice echivalează cu testarea unor astfel de explicații noncauzale. Criteriile Bradford-Hill , discutate în secțiunea 2.1, sunt utilizate pe scară largă în astfel de interpretări.

Utilizarea criteriilor cauzale

Încă de la început, standardele de evidență epidemiologică, conform criteriilor Bradford Hill, trebuie să facă față multor rezerve și excepții. Hill însuși a întrebat: „În ce circumstanțe putem trece de la această asociere observată la un verdict de cauzalitate?" De fapt, el nu a fost de acord că există „reguli de probă stricte și rapide" după care să se judece cauzalitatea (Rothman & Greenland, 2005). Această concluzie este în acord cu multe altele că *inferențe cauzale nu pot atinge certitudinea deducțiilor logice* .

Deși unii oameni de știință continuă să promoveze criteriile cauzale ca ajutoare pentru deducere , alții susțin că este de fapt dăunător să confundăm procesul inferențial prin luarea în considerare a unei liste de criterii. Poate că se poate găsi o abordare intermediară care transformă criteriile în teste deductive ale ipotezelor cauzale. Dacă ar fi așa, s-ar putea evita utilizarea criteriilor cauzale pentru a susține ipoteze și teorii nedemonstrabile, permițând în același timp epidemiologilor să se concentreze pe evaluarea teoriilor cauzale concurente folosind dovezi observaționale critice.

Judecarea dovezilor științifice

Deoarece criteriile cauzale nu pot fi utilizate pentru a stabili validitatea unei inferențe, niciun criteriu nu poate fi utilizat pentru a evalua validitatea datelor ca dovadă. Cu toate acestea, există metode prin care poate fi evaluată validitatea. În acest spirit, s-ar putea privi dovezile științifice ca pe o formă de măsurare.

■ **Exemplu:** Dacă un studiu epidemiologic propune să se determine relația dintre expunerea la fumatul de țigară și riscurile de cancer pulmonar, rezultatele ar trebui privite ca o măsură a efectului cauzal, cum ar fi raportul dintre riscul de cancer pulmonar în rândul fumătorilor și riscul în rândul fumătorilor. nefumători:

Raportul de risc = (Riscul de cancer pulmonar în rândul fumătorilor de țigări)/(Riscul în rândul nefumătorilor)

Măsurarea unui efect cauzal este supusă erorilor de măsurare, la fel ca orice experiment. Pe lângă eroarea statistică, eroarea de măsurare poate include probleme legate de proiectarea studiului, cum ar fi selecția și reținerea subiectului, colectarea de informații , confuzie necontrolată și alte surse de părtinire. Nu este suficient să se caracterizeze un studiu ca

având sau neavând oricare dintre aceste surse de eroare, deoarece aproape fiecare studiu va fi supus aproape fiecărui tip de eroare. Cu toate acestea, validitatea unui studiu trebuie evaluată. Prin urmare, trebuie cuantificate erorile. Deoarece nu există un standard precis cu privire la cât de multă eroare poate fi tolerată înainte ca un studiu să fie considerat invalid, nu există nicio alternativă la cuantificarea erorilor de studiu în măsura posibilului.

Ceea ce se cere este mult mai mult decât aplicarea unei liste de criterii. Trebuie aplicat evaluarea critică pentru a obține o evaluare cuantificată a erorii totale inerente studiului. Acest tip de evaluare necesită abilități cantitative în consonanță cu rigoarea științifică. Se așteaptă ca investigațiile epidemiologice să fie abordate prin aplicarea unor abordări biostatistice viguroase din punct de vedere matematic, așa cum este descris în secțiunea următoare și în restul acestei cărți.

Întrebări de revizuire pentru Secțiunea 2.2

1. În investigațiile EPDM, ori de câte ori se constată o boală, ce presupuneri se fac de obicei cu privire la originea acelei boli? Care sunt bazele acestor presupuneri?
2. (a) Care este o cauză suficientă a unei boli? Da un exemplu.
   (b) Având în vedere o boală, cum determină o investigație epidemiologică cauza suficientă a acesteia?
3. (a) Descrieți diagramele lui Rothman pentru cauze suficiente ale bolilor.
   (b) Cum ilustrează aceste diagrame principiile referitoare la inferența și cauzalitatea bolilor umane?
4. (a) Poate fi adevărat că „nu trebuie să identificăm toate cauzele componente pentru a preveni unele cazuri de boală”? De ce sau de ce nu?
   (b) Dați un exemplu al acestei afirmații.
5. În investigațiile epidemiologice, ce se înțelege prin următoarele?
   (a) Puterea unei cauze
   (b) Interacțiunea între cauze
   (c) Însumarea fracțiilor atribuibile ale cauzelor bolilor
   Dați un exemplu pentru fiecare dintre aceste concepte.
6. (a) În EPDM al studiilor bolilor, care sunt inferențe cauzale?
   (b) În contextul căutării cauzei unei boli, explicați dilema având în vedere imposibilitatea dovezii.
7. Cum se confruntă un epidemiolog sarcina de a testa teoriile epidemiologice concurente? Explicați folosind un exemplu.
8. În investigațiile bolilor epidemiologice, ce se înțelege prin procesul de inferență cauzală?
9. În investigațiile epidemiologice, criteriile Bradford-Hill ajută la confruntarea cu problemele inferenței cauzale? De ce sau de ce nu?
10. În cercetarea epidemiologică, dacă criteriile cauzale nu pot fi utilizate pentru a stabili validitatea unei inferențe, pe ce bază ar trebui să se judece dovezile științifice? De ce?

2.3 BAZELE BIOSTATISTICE DE INFERENȚĂ

Abordarea epidemiologică pentru înțelegerea cauzei bolilor urmează o cale de cauzalitate și inferență, așa cum sa discutat în Secțiunea 2.2. Pentru a susține această abordare, trebuie acum să ne adâncim mai mult în logica inferenței, pentru care o bază biostatistică este alegerea evidentă și logică.

Moduri de inferență

Au apărut și s-au stabilit diferite școli de inferență biostatistică. Aceste școli, sau paradigme, nu se exclud reciproc; metodele care funcționează în mod obișnuit sub o singură paradigmă oferă adesea interpretări atractive sub alte paradigme. Cele două paradigme principale utilizate astăzi sunt biostatistica bayesiană (BIOS) și BIOS-ul frecventist.

În **inferența bayesiană** , teorema lui Bayes este utilizată pentru a calcula modul în care gradul de sprijin pentru (sau credința în) o propoziție se schimbă în funcție de dovezi. Inferența Bayesiană se bazează pe filozofia probabilității bayesiene, care afirmă că gradele de credință (suport) pot fi reprezentate prin probabilități. O aplicație tipică a BIOS-ului bayesian este testarea unei ipoteze.

Inferența frecventistă este o metodă biostatistică alternativă în care se folosește un **test de semnificație pentru a ajunge la concluzia că o propoziție este fie adevărată, fie falsă, sau trage concluzia că un interval de încredere (IC)** dat acoperă valoarea adevărată. În acest cadru, oricare dintre aceste concluzii are o probabilitate dată de a fi corectă, iar această probabilitate are fie o interpretare *a probabilității de frecvență , fie o interpretare pre-experiment* . Un exemplu tipic de aplicare a BIOS-ului frecventist este testarea semnificației unei valori măsurate în ceea ce privește nivelul acesteia într-un CI calculat.

Niveluri de măsurare

Investigațiile biostatistice folosesc patru niveluri de măsurare: raport, interval, ordinal și nominal. Fiecare are grade diferite de utilitate în cercetare.

- **raportului** au o valoare specifică zero. În plus, sunt definite diferențele dintre măsurători, permițând flexibilitate în metodele biostatistice care pot fi utilizate pentru analiza datelor.

- **pe intervale** arată distanțe semnificative dintre măsurători, valoarea zero fiind arbitrară (ca și în cazul măsurătorilor de temperatură folosind scara Celsius sau Fahrenheit).
- **ordinale** au diferențe imprecise sau inegale între valorile consecutive , dar au o ordine semnificativă a acestor valori.
- **nominale** nu au o ordine de rang semnificativă între valori.

Deoarece variabilele aplicabile numai măsurătorilor nominale sau ordinale nu pot fi măsurate numeric, ele sunt uneori grupate ca categorii sau variabile categoriale. În schimb, datorită caracteristicilor lor numerice, măsurătorile de raport și interval sunt grupate împreună ca variabile cantitative, care pot fi fie discrete, fie continue.

BIOS frecventist în EPDM

ESTIMARE INTERVAL

În majoritatea investigațiilor epidemiologice, eșantionul este doar o parte a unei populații; astfel , rezultatele pot să nu fie pe deplin reprezentative pentru întreaga populație, iar estimările obținute din eșantion pot doar aproxima valoarea populației. **CI** le permit biostatisticienilor să exprime cât de aproape se potrivește estimarea eșantionului cu valoarea reală în întreaga populație.

Adesea, acest parametru este setat la **95% CI** . În mod oficial, un IC de 95% pentru o valoare este un interval în care, dacă eșantionarea și analiza ar fi repetate în aceleași condiții (producând un set de date diferit), intervalul ar include valoarea adevărată (populație) în 95% din timp. Aceasta *nu* înseamnă că probabilitatea ca valoarea adevărată să fie în CI este de 95%. Din punct de vedere frecventist, o astfel de afirmație nici măcar nu are sens, deoarece valoarea adevărată nu este o variabilă aleatorie. Fie valoarea adevărată este sau nu în intervalul dat. Cu toate acestea, este adevărat că înainte ca orice date să fie eșantionate și să li se ofere un plan pentru modul în care va fi construit CI, probabilitatea este de 95% ca intervalul care urmează să fie calculat să acopere valoarea adevărată. În acest moment, limitele intervalului sunt variabile aleatoare care nu au fost încă observate. O abordare care dă un interval care poate fi interpretat ca având o probabilitate dată de a conține valoarea adevărată este utilizarea unui interval credibil din statisticile bayesiene. Această abordare depinde de un mod diferit de interpretare a ceea ce se înțelege prin „probabilitate” ; adică ca probabilitate bayesiană.

Rețineți că în BIOS-ul bayesian, un *interval credibil* este un interval din domeniul unei distribuții posterioare de probabilitate utilizat pentru estimarea intervalului (Paoli, Haggard și Shah, 2002) . Generalizarea la probleme multivariate este *regiunea credibilă* . Intervalele Bayes ian credibile sunt analoge cu CI din BIOS-ul frecventist.

De exemplu, într-o investigație epidemiologică care determină distribuția incertitudinii unui parametru clinic l , dacă probabilitatea ca l să fie între 67 și 103 este de 95%, atunci $67 \leq l \leq 103$ este un interval Bayesian credibil de 95% l . Unii scriitori numesc acest lucru „IC Bayesian”, dar în această carte, utilizarea termenului *CI* este limitată doar la intervalul de încredere frecventist, pentru a evita confuzia.

CONCEPTUL DE IMPORTANȚĂ BIOSTATISTICĂ

Biostatistica (BIOS) oferă rareori un simplu răspuns da/nu la întrebarea pusă. Interpretarea se reduce adesea la nivelul de semnificație statistică aplicată numerelor și se referă adesea la probabilitatea (valoarea p) ca o valoare să respingă cu acuratețe ipoteza nulă.

O constatare a semnificației *statistice* nu înseamnă neapărat că rezultatul general este semnificativ în termeni reali. De exemplu, un studiu amplu de droguri poate arăta că medicamentul are un efect benefic biostatistic semnificativ, dar foarte mic, astfel încât este puțin probabil ca medicamentul să ajute pacienții într-un mod vizibil.

Intervalele de încredere în epidemiologie și sănătate publică[8]

Atunci când epidemiologii și practicienii PH folosesc BIOS-ul de sănătate, aceștia pot fi

[8] Interval Bayesian credibil și CI frecventist (disponibil la http://en.wikipedia.org/wiki/Credible interval).

interesați de numărul real de evenimente de sănătate, dar mai des folosesc BIOS-ul pentru a evalua riscul real al unei probleme de sănătate în comunitate. BIOS-ul sănătății observate (acele procente, rate sau numărări care sunt calculate sau estimate din sondaje de sănătate , registre vitale BIOS sau alte sisteme de supraveghere a sănătății) nu sunt întotdeauna o reflectare exactă a riscului real care stau la baza populației. Ratele observate pot varia de la un eșantion la altul sau de la un an la altul, chiar și atunci când adevăratul risc subiacent rămâne același.

BIOS-ul bazat pe eșantioane dintr-o populație este supus unei erori de eșantionare. *Eroarea de eșantionare* se referă la variația aleatoare care are loc deoarece numai subseturi specifice ale întregii populații sunt eșantionate și utilizate pentru a estima o constatare pentru întreaga populație. Este adesea numit în mod eronat „marja de eroare". Chiar și evenimentele de sănătate care se bazează pe o numărare completă a unei întregi populații, cum ar fi decesele, sunt supuse unor variații aleatorii, deoarece numărul de evenimente care au avut loc efectiv poate fi considerat ca unul dintr-o serie mare de rezultate posibile care ar fi putut avea loc în cadrul aceleasi circumstante. În general, eroarea de eșantionare sau variația aleatorie crește atunci când eșantionul, populația sau numărul de evenimente este mic.

Teoria eșantionării statistice poate fi utilizată pentru a calcula un CI pentru a oferi o estimare a discrepanței potențiale dintre parametrii adevărați ai populației și ratele observate. Înțelegerea mărimii potențiale a acestei discrepanțe poate oferi informații despre modul de interpretare a statisticii observate.

Un IC de 95% indică intervalul de valori în care o biostatistică s-ar încadra 95% din timp dacă investigatorul ar calcula biostatistica (de exemplu, o vârstă sau o rată procentuală) dintr-un număr infinit de probe de aceeași dimensiune, extrase din aceeași populație. Cu alte cuvinte, CI este un interval de valori în care se așteaptă să apară valoarea „adevărată" a ratei (cu 95% probabilitate).

Acum, să luăm în considerare cele mai comune metode de calcul al CI de 95% pentru unele rate și estimări utilizate în mod obișnuit în EPDM și PH.

95% CI PENTRU UN PROCENT DIN UN EȘANȚĂ DE SONDAJ

Pentru a calcula un CI pentru un procent dintr-un eșantion de sondaj, trebuie mai întâi să calculați *eroarea standard* a procentului. Un procent este cunoscut și ca medie a unei distribuții binomiale. Eroarea standard a mediei este o măsură a dispersiei pentru distribuția ipotetică a mediilor numită *distribuția de eșantionare a mediei* . Aceasta este o distribuție a mediilor calculate dintr-un număr infinit de eșantioane de aceeași dimensiune extrase din aceeași populație ca eșantionul original. Distribuția *de eșantionare a mediei* are o formă care este aproape identică cu distribuția normală (a se vedea figura 2.3).

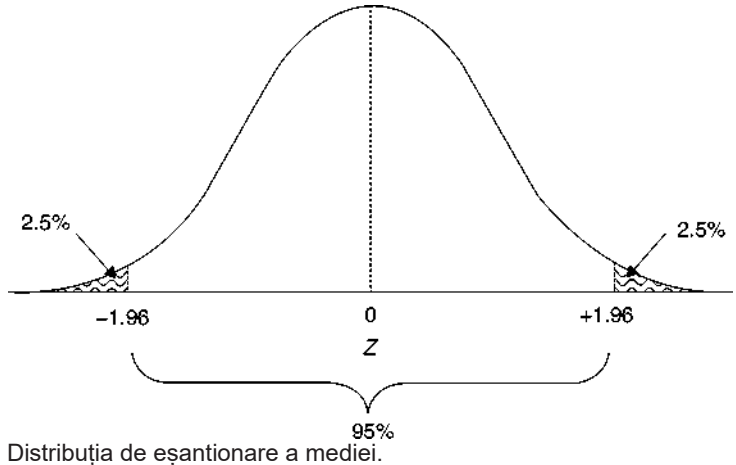


FIGURA 2.3 Distribuția de eșantionare a mediei.

După ce eroarea standard a procentului a fost calculată, atunci se decide cât de mare ar trebui să fie CI. Cea mai frecventă alegere este un IC de 95%. Aceasta este lățimea intervalului care include media (*distribuția de eșantionare a mediei*) 95% din timp. Cu alte cuvinte, un CI de 95% pentru un procent este intervalul de valori în care procentul ar fi găsit cel puțin 95% din timp dacă s-ar obține un eșantion diferit de aceeași dimensiune din aceeași populație. Transformarea erorii standard într-un CI de 95% este simplă.

O *distribuție* este un instrument folosit în BIOS pentru a asocia o biostatistică (de exemplu, un procent, o medie sau o altă statistică) cu probabilitatea acesteia. Când anchetatorii epidemiologici se referă la o măsură ca fiind „semnificativă din punct de vedere biostatistic", ei au folosit o distribuție pentru a evalua probabilitatea biostatisticii și au descoperit că ar fi improbabilă în condiții obișnuite. În cele mai multe cazuri, se poate baza pe măsuri precum rate, medii și proporții pentru a avea o *distribuție normală subiacentă* , cel puțin atunci când dimensiunea eșantionului este suficient de mare.

Trebuie doar să înmulțim eroarea standard cu scorul *z* al punctelor din distribuția normală care exclud 2,5% din distribuția la fiecare capăt (cu două cozi): acel scor *z* este 1,96.

- Un scor *z* de 1,96 definește CI de 95%.
- Un scor *z* de 1,65 definește CI de 90%.

Pentru un eșantion aleator simplu, eroarea standard = $\sqrt{}$ (pq / n) (2.3-1)

unde: p este rata, $q = 1 - p$ și n este dimensiunea eșantionului.

■ **Exemplul A:** Într-un sondaj, 15% dintre respondenți au indicat că au fumat țigări . Eșantionul a fost format din 1.000 de persoane dintr-un eșantion aleator simplu.

Eroarea standard = $\sqrt{}$ [0,15 × (1 - 0,15)/1000] = .0113
CI de 95% este astfel:
0,15 ± 1,96 × eroare standard = 0,15 ± 1,96 × 0,0113
= 0,13 ± 0,0221
= 0,1079 0,1521,

care este adesea afirmat ca (0,1079, 0,1521). Astfel, IC de 95% are o limită inferioară de 10,79% și o limită superioară de 15,21%.

Formula folosită în exemplul precedent se aplică unei distribuții binomiale, care este distribuția a două valori complementare (de exemplu, + și - , 0 și 1, capete și cozi, pro și contra etc.). Pentru a calcula un CI pentru o altă biostatistică, cum ar fi o medie, trebuie să modificați ecuația (2.3-1) . Mărimea *pq* este varianța unei distribuții binomiale.

Dacă măsura nu este o proporție, ci, să zicem, o medie, trebuie să modificați formula, înlocuind varianța cu cantitatea *pq* .

Eroarea standard poate fi calculată și ca abaterea standard împărțită la rădăcina pătrată a dimensiunii eșantionului:

$$\text{Eroare standard} = s / \sqrt{n} \qquad (2.3\text{-}2)$$

PROBE MICI

Dacă eșantionul din care a fost calculat procentul a fost mic [conform teoremei limitei centrale (CLT), se poate defini *mic* ca fiind 29 sau mai puțin; vezi Secțiunea 2.4] , atunci forma distribuției de eșantionare a mediei *nu va* fi aceeași cu forma distribuției normale. Pentru acest caz, se poate folosi o altă distribuție, cunoscută sub numele de distribuție *t* , care are o formă ușor diferită de distribuția normală ; vezi Figura 2.4 .

Procedurile în acest caz sunt similare cu cele discutate pentru CI, dar scorul *t* provine dintr-o familie de distribuții care depind de *gradele de libertate (DF)* . Numărul de DF este definit ca *n* - 1, unde *n* este dimensiunea eșantionului. Pentru o dimensiune a eșantionului de 30, DF este egal cu (30 – 1 = 29). Deci, pentru un CI de 95%, trebuie să utilizați scorul *t* asociat cu 29 de DF. Scorul *t* special este 2,045. Prin urmare, s-ar înmulți eroarea standard cu 2,045 (în loc de 1,96) pentru a genera CI de 95%.

Dacă eșantionul ar avea o dimensiune diferită, să spunem 20, atunci DF ar fi (20 – 1 = 19), care este asociat cu un scor *t* de 2,093 pentru un CI de 95%. Intervalul se va lărgi pe măsură ce

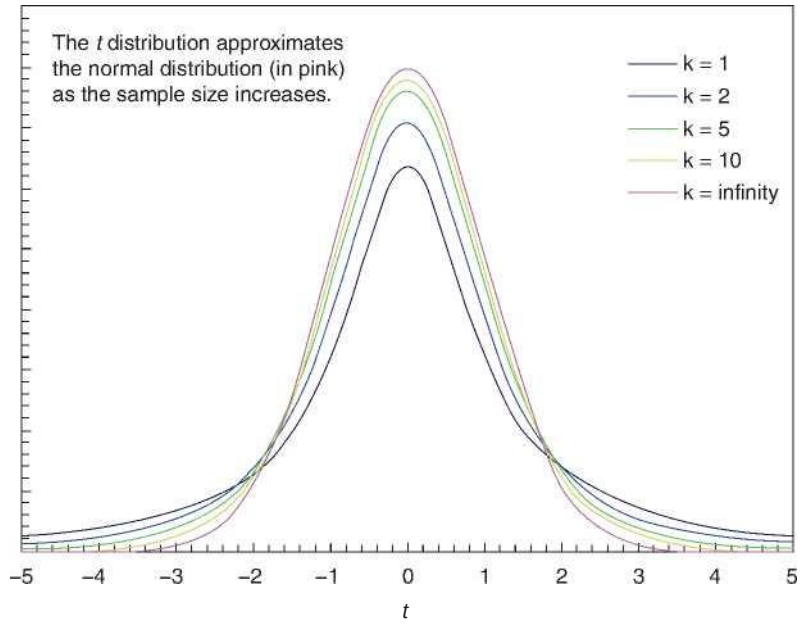


FIGURE 2.4 The Student's *t* distribution at varying DFs (k). The *t* distribution approximates the normal distribution (pink) as the sample size increases.

dimensiunea eșantionului este redusă. Aceasta reflectă incertitudinea în estimarea variației în populație. Pentru un IC de 95% cu 9 DF, scorul *t* este 2,262.

DISTRIBUȚIA T A ELEVULUI LA DIVERSE DF K

Pentru un IC de 95%, s-ar folosi scorul *t* care definește punctele distribuției astfel încât să excludă cele mai extreme 5% din distribuție, care este 0,025 la fiecare capăt al distribuției simetrice.

POPULAȚII FINITE

Atunci când sondajul eșantionează întreaga populație sau cea mai mare parte a populației, utilizarea factorului de corecție al populației finite F va îmbunătăți (scădea) lățimea CI. Acum,

$$F = 1 - f \tag{2,3-3}$$

unde *f* este fracția de eșantionare și

$$f = n / N \tag{2,3-4}$$

unde *n* este dimensiunea eșantionului și *N* este dimensiunea populației.

Fracția de eșantionare *f* este proporția populației care a fost inclusă în eșantion. Eroarea standard (SE) a mediei pentru o distribuție binomială a unui eșantion finit este

$$\text{Procent SE} = \sqrt{\{(\, pq / n\,)\,(1 - f\,)\}} \tag{2.3-5}$$

Aceste tehnici pot fi realizate în mod convenabil utilizând software-ul de biostatistică standard . Secțiunea 2.4 introduce utilizarea mediului de programare R pentru aceste calcule. Sunt disponibile și programe în SAS, SPSS, Stata și alte domenii (deși nu sunt discutate în

acest text).

CÂND RATA ESTE EGALĂ ZERO

Când procentul sau rata este egală cu zero, folosind formulele convenționale se va obține un CI incorect de 0. Pentru a estima CI atunci când procentul sau rata este 0, presupunem doar că numărul de pe numărătorul ratei este 2 și apoi calculați valoarea CI utilizând dimensiunea populației în calculul original.

CI 95% PENTRU EVENIMENTE RARE

Pentru evenimentele rare care apar aleatoriu într-un interval de timp, distribuția normală nu se mai aplică. În astfel de cazuri, distribuția Poisson este utilizată pentru a modela evenimentele, cum ar fi o inundație de 100 de ani. Distribuția Poisson este, de asemenea, utilizată pentru a calcula CI pentru evenimente de sănătate, cum ar fi mortalitatea infantilă sau decesele cauzate de cancer. Această distribuție nu este simetrică față de medie și, prin urmare, CI-urile asociate nu vor fi simetrice (limita superioară este mai departe de estimare decât limita inferioară IS).

Distribuția Poisson ia forma unei distribuții normale atunci când există 20 sau mai multe evenimente în numărător. Prin urmare, utilizați o distribuție Poisson pentru evenimente rare (adică, când numărul de evenimente este mai mic de 20), dar utilizați distribuția normală atunci când numărul de evenimente este de 20 sau mai mult.

DISTRIBUȚIE OTRAVĂ

Pentru a calcula CI folosind o distribuție Poisson, înmulțiți rata estimată cu factorul de încredere asociat cu numărul de evenimente pe care se bazează rata.

■ **Exemplul B:** Într-o anumită zonă geografică, au existat 853 de nașteri într- un singur an și 8 decese infantile în același an. Prin urmare, rata mortalității infantile a fost de 9,4 la 1.000 de născuți vii, calculată ca [(8/853) × 1.000]. Limitele inferioare și superioare de încredere sunt calculate utilizând factorii de încredere găsiți în Anexa B a instrumentului auxiliar online. Factorii pentru șapte evenimente sunt 0,4317 și 1,9704 pentru limitele inferioare și, respectiv, superioare ale CI.

Prin urmare,

limita inferioară a CI = 9,4 × 0,4317 = 4,05 și

limita superioară a CI = 9,4 × 1,9704 = 18,52

pentru o rată a mortalității infantile de 9,7 la 1.000 de născuți vii, cu un IC de 95% de la 4,05 la 18,52.

RATA DE DECES AJUSTATĂ DIRECT ÎN VÂRSTE (DAADR)

Când se compară între zone geografice, o anumită metodă de ajustare a vârstei este de obicei utilizată pentru a controla diferențele de la o zonă la alta în ceea ce privește evenimentele de sănătate, care pot fi explicate prin vârstele diferite ale populațiilor din zonă. De exemplu, o zonă care are

> o populație mai în vârstă va avea rate brute mai mari (neajustate în funcție de vârstă) pentru cancer, chiar dacă nivelurile sale de expunere și ratele de cancer pentru anumite grupuri de vârstă sunt aceleași cu cele din alte zone. S-ar putea atribui incorect ratele mari de cancer unor caracteristici ale zonei, altele decât vârsta. Ratele ajustate în funcție

de vârstă controlează efectele vârstei, permițând o mai bună comparabilitate a ratelor între zone. Standardizarea directă ajustează ratele specifice vârstei observate în zona mică la distribuția de vârstă a unei populații standard. Rata de deces direct ajustată în funcție de vârstă (DAADR) este o medie ponderată a ratelor de deces specifice vârstei, unde ponderile specifice vârstei reprezintă distribuția relativă de vârstă a populației standard:

Rata mortalității ajustată direct în funcție de vârstă (DAADR) = $\Sigma W_{si} \times D_{.}/P_{.} = \Sigma W_{si} \times R_{i}$

unde W_{si} = ponderea pentru a *i* -a grupă de vârstă în populația standard (proporția - populației standard în a *i* -a grupă de vârstă):

$$= P_{si} / \Sigma P_{Si}$$

P_{si} = populația din grupa de vârstă *i* din populația standard
D_{i} = numărul de decese (sau alt eveniment) în grupa de vârstă *i* a populației studiate
P_{i} = populația din grupa de vârstă *i* din populația studiată
R_{i} = rata specifică vârstei în grupa *a i* -a de vârstă.

Folosind proprietățile distribuției Poisson, varianța ratei mortalității specifice vârstei este dată de

$$\text{var}(R_{i}) = \text{var}(D_{i}/P_{i}) = 1/P_{i}^{2}\,\text{var}(D_{i}) = D_{i}/P_{i}^{2} = R_{i}^{2}/D_{i}$$

Varianta unui DAADR poate fi calculată după cum urmează:

$$\text{var(DAADR)} = \sum W_{si}^{2} \times \text{var}(R_{i}) = \Sigma W_{si}^{2} \times R_{i}^{2}/D_{.}$$

$$\text{SE(DAADR)} = \sqrt{\{\text{var(DAADR)}\}}$$

unde:

var(DAADR) = varianța ratei standardizate direct
W_{si} = ponderea celui de -*al i* -lea grup de vârstă în populația standard
R_{i} = rata specifică vârstei în grupa *a i* -a de vârstă
$\text{var}(R_{i})$ = varianța ratei mortalității specifice vârstei în al *i* -lea grup de vârstă al populației studiate

$$= R_{i}^{2}/D_{i}$$

D *i* = numărul de decese (sau alte evenimente) în grupa de vârstă *i* a populației studiate
SE(DAADR) = eroarea standard a ratei standardizate direct.

Rata de deces ajustată în funcție de vârstă este o combinație liniară de variabile aleatoare Poisson independente și, prin urmare, nu este ea însăși o variabilă aleatoare Poisson. Poate fi plasat în familia mai generală a distribuțiilor gamma din care face parte Poisson.

Majoritatea pachetelor de software biostatistice au o funcție de calculare a factorilor care pot fi aplicați ratelor de deces ajustate în funcție de vârstă pentru a calcula CI de 95%. Acești factori sunt derivați dintr-o distribuție gamma standard.

TARIFE AJUSTATE INDIRECT PE VÂRĂ

Metoda directă poate prezenta probleme atunci când dimensiunile populației sunt deosebit de mici. Calculul direct al ratelor standardizate necesită calcularea ratelor specifice vârstei, iar pentru zone mici, aceste rate specifice vârstei se pot baza doar pe unul sau două evenimente. Regula generală este că, dacă există mai puțin de 20 până la 25 de cazuri în populația indexată, ar trebui utilizată standardizarea indirectă a ratelor.

Ratele standardizate indirect se bazează pe rata de mortalitate standardizată (SMR) și pe rata brută pentru o populație standard. Standardizarea indirectă ajustează rata generală a populației standard la distribuția de vârstă a zonei mici. Este valabil să se compare ratele standardizate indirect doar cu rata din populația standard; nu pot fi comparate între ele.

O rată de deces sau de boală (ISR) standardizată indirect poate fi calculată ca

$$ISR = SMR*Rs$$

SMR = decese/boală observate în zona mică = D/e = D/Σ ($R_{si} \times n_i$) decese/boli preconizate în zona mică, unde:

SMR = decese observate în zona mică/decese așteptate în zona mică

D = numărul de decese observat în zona mică

e = $\sum (R_{s\,i} \times n.$) = numărul așteptat de decese în zona mică

R_s = rata brută a mortalității în populația standard

R_{si} = rata de deces specifică vârstei în grupa de vârstă *i* a populației standard = # decese/număr de populație, înainte de a aplica constanta

n_i = numărul populației din grupa de vârstă *i* a zonei mici.

Când raportul dintre evenimente și populația totală este mic (<0,3) și dimensiunea eșantionului este mare, următoarele două metode pot fi utilizate pentru a calcula CI.

(1) Când numărul de evenimente ≥ 20:

$$CI_{ISR} = \pm 1{,}96 \sqrt{(SMR/e)} \times R_s \times K$$

unde:

SMR = decese observate în zona mică/decese așteptate în zona mică e = decese așteptate în zona mică = Σ ($R_{si} \times n_i$)

R_s = rata brută a mortalității în populația standard

R_{si} = rata mortalității specifice vârstei în grupa de vârstă *i* a populației standard (# decese/număr de populație)

n_i = numărul populației din grupa de vârstă *i* a zonei mici

K = o constantă (de exemplu, 100.000) care este utilizată pentru a comunica rata.

(2) Când numărul de evenimente ≤ 20:

estimarea parametrilor din tabelul Poisson /e) × R_s × K ULISR = (limita superioară pentru estimarea parametrilor din tabelul Poisson /e) × R_s × K unde LL este limita inferioară CI și UL este limita superioară CI.

Interval Bayesian credibil [3]

În statistica bayesiană, un **interval credibil** (sau intervalul credibil bayesian) este un interval din domeniul unei distribuții posterioare de probabilitate utilizat pentru estimarea

dimensiunea unui interval. Generalizarea la probleme multivariate este **regiunea credibilă** . Intervalele credibile sunt analoge cu CI din BIOS-ul frecventist. De exemplu, într-un experiment care determină distribuția incertitudinii parametrului t , dacă probabilitatea ca t să fie între 35 și 45 este de 90%, atunci $34 \leq t \leq 45$ este un interval credibil de 90%.

UN INTERVAL BAYESIAN CREDIBIL VERSUS UN CI FRECVENTIST

Un IC 90% frecventist de 35 până la 45 înseamnă că, cu un număr mare de mostre repetate , 90% din IC calculate ar include valoarea reală a parametrului. Probabilitatea ca parametrul să se afle în intervalul dat (să zicem, 35–45) este fie 0, fie 1 (parametrul necunoscut nealeatoriu este fie acolo, fie nu). În termeni frecventiști, parametrul este *fix* (nu poate fi considerat a avea o distribuție a valorilor posibile) iar CI este *aleatoriu* (deoarece depinde de eșantionul aleatoriu). Astfel, un CI este un interval generat de o procedură care va da intervale corecte în 90% din timp.

În general, intervalele Bayesiene credibile nu coincid cu CI-urile frecventiste, din două motive:

1. Intervalele credibile încorporează informații contextuale specifice problemei din distribuția anterioară, în timp ce CI se bazează doar pe date.
2. Intervalele credibile și CI tratează parametrii deranjanți în moduri radical diferite.

Mulți biostatisticieni profesioniști și oameni de știință în materie de decizie, precum și specialiștii non-statistici, interpretează intuitiv CI în sensul intervalului credibil bayesian și, astfel, intervalele credibile sunt uneori numite și CI. Este larg acceptat, în special în științele deciziei, că intervalul credibil este doar subsetul subiectiv al CI. De fapt, multe cercetări în evaluările calibrate ale probabilității nu utilizează niciodată termenul *interval credibil* ; este obișnuit să folosiți pur și simplu *CI* .

Întrebări de revizuire pentru Secțiunea 2.3

1. (a) Care sunt cele două moduri principale de inferențe statistice utilizate în BIOS?
   (b) Care sunt asemănările și diferențele dintre ele? Dați exemple.
2. (a) În analiza datelor BIOS, ce este un CI?
   (b) Ce înseamnă o afirmație de genul „Sunt 95% încrezător că X este între (acesta) și (aceasta)”?
3. Ca investigator epidemiologic, cum ar trebui să răspundem la următoarele întrebări?
   (a) Dacă rata fumatului în rândul adolescenților scade de la 12% la 10%, ar trebui să sărbătorim asta ca o scădere semnificativă? De ce sau de ce nu?
   (b) Dacă rata mortalității infantile de stat de 6,75 crește la 7,05 într-un interval de 1 an, ar trebui să ne îngrijorăm această creștere? De ce sau de ce nu?
4. (a) Care este distribuția normală?
   (b) *t* a Studentului ?
   (c) Cum sunt legate aceste două?
5. Explicați ce se înțelege prin „intervale de încredere în epidemiologie și sănătate publică” și modul în care BIOS aplică CI în analiza epidemiologică.

2.4 BIOS ÎN EPDM ȘI PH

Prin definiție, *biostatistica (BIOS)* este aplicarea statisticii la un subiect din biologie. În medicină și științele sănătății, aceste aplicații includ proiectarea de experimente biologice, colectarea și analiza datelor din acele experimente și interpretarea rezultatelor. Biostatisticienii care comunică profesional cu privire la activitățile, munca și rezultatele lor pot ajuta la construirea unei aprecieri mai largi a BIOS-ului.

Aplicații ale BIOS-ului[9]

BIOS-ul găsește de obicei aplicație în:

- PH, inclusiv EPDM, cercetarea serviciilor de sănătate, nutriția și sănătatea mediului.
- Proiectarea și analiza studiilor clinice în medicină.
- populației și genetica statistică, adesea pentru a lega o variație a genotipului cu o variație a fenotipului. Acest lucru a fost făcut în agricultură pentru a îmbunătăți culturile și animalele de fermă (prin creșterea selectivă a animalelor) . În cercetarea biomedicală, această muncă poate ajuta la găsirea de candidați pentru alelele genelor care pot cauza sau influența predispoziția la boli în genetica umană.
- Analiza datelor genomice (de exemplu, din experimente cu microarray sau proteomică , care se referă adesea la boli sau stadiile bolii).
- Ecologie și prognoză ecologică.
- secvenței biologice .
- sistemelor , pentru inferența rețelelor de gene sau analiza căilor .

Metodele statistice încep să fie integrate în informatica medicală, informatica PH, bioinformatica și biologia computațională. Munca profesională în aceste domenii este adesea publicată în **reviste BIOS,** cum ar fi următoarele:

Biometrie
Biometrica
Biostatistica
Jurnalul canadian de epidemiologie și biostatistică
Jurnalul Internațional de Biostatistică
Jurnalul de Statistică Agricolă, Biologică și de Mediu
Jurnalul de biometrie și biostatistică
Jurnalul de Statistică Biofarmaceutică
Statistica farmaceutică
Aplicații statistice în genetică și biologie moleculară
Statistica în cercetarea biofarmaceutică
Statistica în Medicină
Turkiye Klinikleri Journal of Biostatistics

BIOS-ul partajează mai multe metode cu câmpuri cantitative, inclusiv:

- Biologie computațională
- Informatică
- Cercetare operațională
- Psihometria
- Statistici
- Econometrie
- Demografie matematică

[9]Biostatistics (preluat de la http://en.wikipedia.org/wiki/Biostatistics).

BIOS în EPDM și PH

În domeniile EPDM și PH, BIOS ocupă o poziție esențială în analiza datelor pentru a oferi concluzii și concluzii utile și semnificative pe care se pot baza deciziile și politicile PH critice și utile. Aceste activități includ:

- Generarea de analize descriptive pentru afișarea și rezumarea datelor
- Aplicarea conceptelor care stau la baza inferenței statistice
- Utilizarea metodelor biostatistice pentru analiza datelor continue și binare
- Aplicarea aspectelor biostatistice ale designului studiului , cum ar fi eșantionarea, distribuțiile de probabilitate și distribuția de eșantionare a mediei; CI și teste de semnificație pentru un eșantion, două eșantioane pereche și două eșantioane independente pentru date continue, precum și date binare
- Căutând corelația și regresia liniară simplă
- Luând în considerare metodele fără distribuție pentru două eșantioane pereche, două eșantioane independente și corelație și estimarea puterii și a mărimii eșantionului pentru studii simple
- Introducerea aspectelor statistice ale proiectării și analizei studiului

OBIECTIVE DE ÎNVĂȚARE (UNIVERSITATEA DIN SYDNEY, 2012)

Elevul care citește și învață cu succes din această carte va fi capabil să:

1. Rezumați datele biostatistice folosind tabele, grafice și BIOS-ul rezumat corespunzător.
2. Interpretați testele de semnificație și CI.
3. *t* Student pentru variabile continue și testul chi-pătrat pentru date categorice, atât în cazurile pereche, cât și în cele nepereche; calculați CI pentru principalele rezultate; și rezumați concluziile din astfel de analize.
4. Comparați două eșantioane folosind teste neparametrice, atât în cazuri pereche, cât și în cazuri nepereche, și rezumați concluziile din astfel de analize.
5. Analizați asocierea dintre două variabile folosind diagrame de dispersie, măsuri parametrice și neparametrice de corelație și regresie liniară simplă.
6. Calculați dimensiunea eșantionului necesară pentru studii simple.
7. Creați un plan de analiză statistică, care detaliază etapele majore în proiectarea și analiza statistică a unui studiu.
8. Utilizați software-ul statistic R pentru a procesa, analiza și prezenta date.

Studenților li se va cere să efectueze analize folosind un calculator și, de asemenea, li se va cere să efectueze analize folosind software-ul statistic R .

APLICAREA BIOS-ului ÎN EPDM (CDC, 2006; VIRASAKDI, 2011; ARAGON, 2011)

Multe variabile folosite de epidemiologi sunt variabile categorice. Unele dintre acestea au doar două categorii:

- Expus: DA sau NU
- Testat: + sau - , caz sau control și așa mai departe

Aceste variabile trebuie rezumate cu măsuri de frecvență, cum ar fi rate, rapoarte și proporții. Trei măsuri de frecvență care sunt utilizate pentru a caracteriza apariția evenimentelor de sănătate într-o populație sunt incidența, prevalența și ratele de

mortalitate.

Pentru început, luăm în considerare aceste concepte prin calcularea și interpretarea următoarelor măsuri epidemiologice:

- Raport
- Proporţie
- Proporția de incidență (rata de atac)
- Rata de incidenta
- Prevalența
- Rata mortalității

Prelucrarea și analiza datelor epidemiologice de bază

Datele de bază în investigațiile epidemiologice pot fi analizate în funcție de următorii parametri și abordări.

MĂSURI DE FRECVENȚĂ

O măsură a locației centrale oferă o singură valoare care rezumă o întreagă distribuție a datelor. În schimb, o măsură de frecvență caracterizează doar o parte a distribuției. Măsurile de frecvență compară o parte a distribuției cu o altă parte a distribuției sau cu întreaga distribuție. Măsurile comune ale frecvenței sunt **rapoartele** , **proporțiile** și **rata s** . Toate cele trei **măsuri de frecvență** au aceeași formă de bază:

$$(\text{Numărător/Numitor}) \times 10^{n}$$

unde n = …, 3, 2, 1, 0, 1, 2, 3, ….

RAPORT

Un **raport** este mărimea relativă a două mărimi sau o comparație a oricăror două valori . Se calculează prin împărțirea unei variabile la scară de interval sau raport la cealaltă. Numătorul și numitorul nu trebuie să fie relaționați. Prin urmare, s-ar putea compara merele cu portocale (sau merele cu numărul de vizite clinice).

Metoda de calculare a unui raport folosește numărul sau rata de evenimente, articole, persoane și așa mai departe dintr-un grup împărțit la numărul sau rata de evenimente, articole, persoane și așa mai departe dintr-un alt grup. După ce numărătorul este împărțit la numitor,

TABELUL 2.2 Măsuri epidemiologice de apariție și asociere

TIP	MĂSURI [a]	DESCRIERE
Măsuri de apariție	Incidența timpului [b]	Timp mediu într-o stare („supraviețuire") Număr de cazuri noi
	Prevalența [c]	Numărul de cazuri existente
	Rata	Cazuri noi per persoană-timp cu risc
	Riscul [d]	Probabilitatea de a deveni un caz
	Cote	Șansele de a deveni un caz
Măsuri de asociere	Rata ratei	Comparația a două rate
	Raportul de risc (RR)	Comparația a două riscuri
	Rata cotelor (OR)	Comparația a două cote

[a] Toate măsurile au un element de timp care trebuie specificat.

[b] Uneori exprimat ca proporție prin împărțirea la „populația totală". [c] Exprimat în mod obișnuit ca proporție prin împărțirea la „populația totală". [d] Estimată uneori folosind o proporție (model binomial).

rezultatul este exprimat sub forma „[număr] la [număr]" (de exemplu, „3 la 1") sau ca „[număr]:[număr]" (de exemplu, „3:1").

Tabelele 2.2 și 2.3 prezintă măsurile de apariție și măsurile de asociere utilizate în mod obișnuit în EPDM. Printre acești parametri, cercetătorii se concentrează mai întâi pe incidență, prevalență, rata, riscul, șansele și raportul dintre rate (raportul ratei), riscurile (raportul de risc [RR]) și șansele (raportul cotelor [OR]). Incidența și ratele sunt determinate de numărul de cazuri noi care apar într-o perioadă de timp. Aceste procese pot fi reprezentate prin modele de probabilitate Poisson.

TABELUL 2.3 Măsuri epidemiologice: rapoarte, proporții și rate

STARE	RAPORT	PROPORŢIE	RATA
Morbiditate (boală)	Raportul de risc (RR; risc relativ) Rata ratei Raportul de cote (OR) Prevalența perioadei	Rata de atac (proporția de incidență) Rata de atac secundar Prevalență punctuală Proporție atribuibilă	Rata de incidență persoană-timp
Mortalitatea (moartea)	Raportul deces-caz	Mortalitate proporțională	Rata brută a mortalității Rata de letalitate Rata mortalității specifice cauzei Rata mortalității specifice vârstei Rata mortalității materne Rata mortalității infantile
Natalitate (naștere)			Crude birth rate Crude fertility rate

TABELUL 2.4 Metode de estimare utilizate în mod obișnuit pentru măsurile epidemiologice

MĂSURĂ	PROBA MARE	PROBA MICĂ

Un singur eșantion	UMLE	MUE
Două mostre	UMLE	MUE, CMLE, SSAE

Abrevieri: CMLE = estimarea maximă de probabilitate condiționată; MUE = estimare mediană imparțială; SSAE = estimarea ajustării eșantionului mic; UMLE = estimare necondiționată a probabilității maxime.

Prevalența și riscul pot fi estimate folosind proporții, reprezentate prin modele de probabilitate binomică. Alegerea metodei de estimare punctuală depinde dacă se estimează o măsură cu unul sau două eșantioane, de dimensiunea eșantionului și de alegerea metodelor (rezumate în Tabelul 2.4).

Raportul de șanse dintr-un tabel de 2 × 2 oferă un exemplu de estimare a probabilității maxime necondiționate (UMLE), estimare a probabilității maxime condiționate (CMLE), estimare mediană imparțială (MUE) și estimare a ajustării eșantionului mic (SSAE). Folosind notația pentru un tabel 2 × 2 (Tabelul 2.5), cota UMLE este raportul dintre șansele de boală pentru un studiu de cohortă sau raportul dintre șansele de expunere pentru un studiu caz-control:

$$\text{Rata de șanse de boală} \equiv \text{DOR} = (A_1/B_1)/(A_0/B_0) = A_1B_0/A_0B_1 \quad (2.4\text{-}1)$$

$$\text{Raportul cotelor de expunere} \equiv \text{EOR} = (A_1/A_0)/(B_1/B_0) = A_1B_0/A_0B_1 \quad (2.4\text{-}2)$$

$$\text{Prin urmare, DOR} = \text{EOR} \quad (2.4\text{-}3)$$

Tabelul 2.6 prezintă date dintr-un studiu de cohortă privind diareea la sugarii alăptați infectați cu o serie de bacterii holera. Apariția diareei a fost evaluată prin compararea sugarilor ale căror afecțiuni sunt indicate de niveluri scăzute față de maxime ale diluțiilor probei de anticorpi detectate.

Raportul de șanse ale bolii UMLE poate fi apoi calculat după cum urmează:

$$\text{SAU}_{\text{UMLE}} = (12/2)/(7/9) = (12)(9)/(7)(2) = 108/14 = 7{,}7143$$

Acest lucru dezvăluie că șansele de a dezvolta diaree au fost de 7,7 ori mai mari la sugarii cu titruuri scăzute de anticorpi, comparativ cu sugarii cu titruri mari de anticorpi.

TABELUL 2.5 Notații pentru un tabel brut 2 × 2

	EXPUS	NEEXPUS	TOTAL
Cazuri	A_1	A_0	M_1
Noncazuri	B_1	B_0	M_0
Total	N_1	N_0	N

TABELUL 2.6 Comparația diareei la 30 de sugari alăptați colonizați cu *Vibrio cholerae* , în funcție de titrurile de anticorpi din laptele matern al mamei

	NIVEL DE ANTICORPI		
UT	SCĂZ	RIDICAT	TOTAL
Diaree	12	7	19
Fără diaree	2	9	11
Total	14	16	30

Calculul CMLE OR se obține folosind distribuția hipergeometrică pentru tabele cu numere mici, cum ar fi Tabelul 2.5 . În acest caz, se pot trata marjele ca fiind fixe și se pot modela distribuția lui A_1. Ecuația hipergeometrică (Ecuația 2.4-4) este baza testului exact al lui Fisher:

$$\Pr(A_1=a_1|m_1,m_0,n_1,n_0)= \frac{\binom{n_1}{k}\binom{n_0}{m_1-k}OR^k}{\sum_k \binom{n_1}{k}\binom{n_0}{m_1-k}OR^k} \qquad (2.4\text{-}4)$$

unde *k* variază peste toate valorile posibile ale lui A_1. Soluția OR în ecuație (2.4-4) este CMLE OR.

În mediul de calcul R , se poate folosi funcția specială fisher.test pentru a calcula CMLE OR cu rezultatul: $OR_{CMLE} = 7{,}17$.

MUE OR se bazează pe calcularea exactă a valorilor *p inferioare și superioare* . Folosind acest exemplu hipergeometric, MUE OR satisface ecuația (2.4-4) când:

inferior superior .

sau

$$0 = P \quad - P \qquad (2.4\text{-}5)$$

Cu ecuația (2.4-5), se pot folosi funcțiile uniroot și fisher.test din R pentru a calcula MUE OR, cu rezultatul: $OR_{MUE} = 6{,}88$.

Mai mult, atât pentru eșantioanele mici, cât și pentru cele mari, se pot folosi funcții simple în R pentru a calcula un OR pentru un tabel brut 2 × 2 (Tabelul 2.7).

TABELUL 2.7 Rezumatul estimării raportului de cote folosind datele din tabelul 2.5

METODĂ	RAPORTUL COTELOR	COMENTARIU
MLE necondiționat	7.7	Eșantion mare
MLE condiționată	7.2	Probă mică
Estimare mediană imparțială	6.9	Probă mică
Eșantion mic ajustat	4.5	Probă mică; zero la numitor

NOTĂ :

(1) Dacă numărul observat sau așteptat în orice celulă < 5, atunci utilizați una dintre metodele de eșantionare mică pentru estimare (Rothman, 1976) .

(2) Dacă există un zero în numitorul unui UMLE, atunci utilizați ajustarea lui Jewell

pentru eșantionul mic sau tratamente matematice similare, deoarece împărțirea la zero este imposibilă (Rothman, 1976).

Analiza datelor epidemiologice

Un buletin de știri de seară a raportat o serie de articole legate de sănătate:

1. Creșterea nivelului de smog într-un cartier din centrul orașului a fost acuzată de creșterea atacurilor de astm în rândul populațiilor de copii și vârstnici.
2. Un tânăr care fusese anterior diagnosticat cu diabet zaharat de tip I (tipul dependent de insulină) a avut un nivel semnificativ îmbunătățit de control al glicemiei după ce a urmat un curs de tratament medical nestandard, folosind medicina tradițională chineză.
3. Recomandările revizuite ale Departamentului Federal de Sănătate pentru cine ar trebui să primească vaccinul antigripal în acest an.
4. Un raport despre strategiile extinse de monitorizare a bolilor implementate într-un oraș din sud, recent afectat de un uragan masiv.
5. O descriere a unei descoperiri publicate recent într-un jurnal medical de top privind o asociere între expunerea la anumite substanțe chimice și un risc crescut de cancer.

Fiecare dintre aceste știri a inclus interviuri cu oficiali PH numiți *epidemiologi* . Munca lor și modul în care se aplică EPDM sunt subiectele principale ale acestei cărți.

Pentru început, investigațiile epidemiologice iau în considerare diverse evenimente legate de sănătate și efectuează diverse măsurători cu scopul de a lega cantitativ cauze și efecte. Pentru a studia orice astfel de relații cauză-efect, se poate începe prin a analiza:

- Măsurători cu un singur eșantion (măsurarea apariției: rata, riscul și prevalența)
- Măsurători din două eșantioane (asocieri de măsurare: raportul ratei, RR și OR)

Pentru a analiza evenimentele, se poate lua în considerare:

1. Estimarea punctuala:
   - Măsurile de apariție sunt măsuri cu un singur eșantion.
   - Măsurile de asociere sunt măsuri din două eșantioane.
2. Variabilitatea și precizia acestei estimări, folosind CI
3. Valoarea de referință, calculând valorile *p* și erorile de tip I
4. Diferențele față de referințe (dimensiunea efectului, puterea și erorile de tip II)
5. Dimensiunile efective ale eșantionului:
   - lățimea intervalului de încredere dorită
   - Pentru a detecta dacă o măsură dintr-un singur eșantion diferă de o valoare de referință
   - Pentru a detecta dacă o măsură din două eșantioane diferă de o valoare de referință

Estimarea punctuală permite inferențe epidemiologice, cum ar fi:

- Riscul de apariție a bolii sau prevalența unei afecțiuni
- RR pentru dezvoltarea bolii în rândul persoanelor expuse comparativ cu persoanele neexpuse

CI oferă informații despre:

- Varianta estimării ; varianța depinde de variabilitatea naturală a măsurii (de exemplu, greutatea și înălțimea la diferite populații)
- aleatorie de măsurare (de exemplu, instrument, tehnician)

- Dimensiunea eșantionului (creșterea dimensiunii eșantionului îmbunătățește precizia estimării și reduce CI)

Adesea, BIOS-ul este utilizat pentru a determina dacă o estimare punctuală este în concordanță cu o valoare de referință. Presupunând că estimarea provine dintr-o distribuție cu media egală cu valoarea de referință (ipoteza nulă), putem calcula probabilitatea (valoarea *p* bifață) de a obține valoarea biostatistică de test sau valori mai extreme . Dacă ipoteza nulă este adevărată, dar este respinsă incorect deoarece valoarea *p* este mai mică decât .05 aleasă în mod arbitrar, atunci există o eroare de tip I.

Dacă estimarea punctuală este „consecventă" ($p > \alpha$), atunci trebuie să ne întrebăm dacă dimensiunea eșantionului a fost suficientă pentru a detecta o diferență semnificativă, dacă a existat una. Ar trebui evitată greșeala de a deduce că nu a existat nicio diferență atunci când dimensiunea eșantionului a fost prea mică pentru a susține o astfel de concluzie.

Acest lucru necesită definirea a ceea ce înțelegem prin *diferență semnificativă* și apoi calcularea probabilității de a detecta dimensiunea efectivă (sau mai mare), dacă aceasta există. Această probabilitate este **puterea biostatistică** ($1 - b$). O dimensiune efectivă implică o ipoteză alternativă. Probabilitatea de a nu detecta dimensiunea efectivă în cadrul ipotezei alternative este desemnată ca b . Este o eroare de tip II.

Dacă se decide să eșantioneze o populație pentru a estima măsurile epidemiologice, dimensiunea necesară a eșantionului depinde dacă biostatistica este o măsură cu un eșantion (măsuri de apariție) sau o măsură cu două eșantioane (măsuri de asociere). Pentru o măsură cu un singur eșantion, se pot folosi următoarele:

- O dimensiune a eșantionului pentru a obține o lățime CI dorită (utilizată pentru studii descriptive).
- O dimensiune a eșantionului pentru testarea ipotezelor (diferență semnificativă față de o anumită valoare de referință). Acest lucru necesită setarea erorilor de tip I (α) și erorilor de tip II (b).

ESTIMARE

Această procedură a fost discutată mai devreme în această secțiune. Caracteristici și -metodologii suplimentare sunt prezentate în discuția despre mediul de calcul R.

INTERVALE DE ÎNCREDERE

Conceptul foarte util al CI a fost discutat în Secțiunea 2.3. Caracteristici și metodologii suplimentare sunt prezentate în discuția despre mediul de calcul R.

TESTARE DE IPOTEZE ŞI VALORI P

O întrebare rezonabilă în evaluarea unei măsuri epidemiologice este: „Cât de compatibilă este estimarea *r* în comparație cu o anumită valoare de referință?"

- **Exemplu:** Dacă *r* proporția de pacienți a suferit o recidivă gravă anul trecut, iar obiectivul viitor este să nu aibă mai mult de r_0 proporție care se confruntă cu o complicație similară, cât de compatibilă este experiența *r* cu o valoare de referință de r_0 ?

Pentru a răspunde la aceasta, se poate specifica ipoteza nulă:

$$r = r_0$$

Pentru o populație eșantion dată, se poate calcula probabilitatea *p* de a observa o valoare de *r* sau mai mare în ipoteza nulă:

$$p = P(r \geq r \mid r = r_0)$$

Deoarece cineva este interesat doar de o *creștere a complicațiilor, o valoare p* unilaterală este adecvată și adecvată. Dacă această valoare *p* este mare, atunci estimarea (*r*) este mai compatibilă cu ipoteza nulă. Acest rezultat înseamnă că valorile lui *r* sau mai extreme sunt probabil să apară doar întâmplător (o eroare aleatorie). Totuși, dacă această valoare *p* este mică, atunci estimarea este mai puțin compatibilă cu ipoteza nulă; adică este puțin probabil ca valorile lui *r* sau mai extreme să apară doar întâmplător. Valorile *p* ridicate (valorile *p* mai compatibile cu ipoteza nulă) pot apărea deoarece ipoteza nulă este adevărată sau pentru că dimensiunea eșantionului este prea mică pentru a detecta o ipoteză alternativă (adică putere statistică insuficientă).

Ca și în cazul CI, pentru a calcula valoarea *p* , trebuie fie să cunoaștem, fie să facem ipoteze despre distribuția relevantă. De exemplu, să presupunem că 5% dintre internările în spital au dus la recăderi grave și scopul a fost să rămână la sau sub o valoare de referință de 2%. Ar trebui să vă asigurați că 5% observat este compatibil cu obiectivul de referință de 2% sau mai mic. Valoarea *p* unilaterală depinde de mărimea diferenței (5% față de 2%) și de numărul de spitalizări:

(a) Calculați probabilitatea binomială a obiectivului de referință de 2%.
(b) Să spunem că 5 din 100 de internări în spital (5%) au dus la o recădere gravă. Sunt 5 recidive compatibile cu obiectivul de referință de 2% sau mai puțin?

$$k = 5$$

$$n = 100$$

$$r = k / n = 5/100 = 0{,}05$$

$$r_0 = 2\% = 0{,}02$$

Probabilitatea binomială necesară este dată de

$$p = P(r \geq r \mid r_0) = P\,(r \geq 0{,}05 \mid r_0 = 0{,}02) \qquad (2{,}4\text{-}6)$$

Pentru a găsi probabilități dintr-o distribuție binomială, le puteți calcula direct sau utilizați un calculator de distribuție binomială. Pentru a evalua compatibilitatea, se poate folosi un computer cu software-ul adecvat - în acest caz, mediul R. Aici, toate cele trei metode vor fi folosite pentru ilustrare.

CALCUL DIRECT. În general, dacă variabila aleatoare *K* urmează distribuția binomială B(*n* , *p*) cu parametrii *n* și *p* , se poate scrie *K* ~ B(*n* , *p*). Probabilitatea de a obține exact *k* succese în *n* încercări este dată de funcția de masă a probabilității

$$f(k;n,p)=\Pr(K=k)= \binom{n}{k} {}^{kn-k}p(1 - \qquad (2.4\text{-}7)$$

for $k = 0, 1, 2, ..., n$, where

$$\binom{n}{k} \qquad \frac{n!}{k!(n-k)!} \qquad (2.4\text{-}8)$$

$$={}_nC_k \qquad (2.4\text{-}9)$$

Prin urmare, folosind un calculator portabil și lucrând cu patru zecimale,

$$\Pr(k = 2) = f(2; 100, 0{,}05) = {}_n\mathrm{C}_k p^k (1-p)^{n-k}$$
$$= {}_{100}\mathrm{C}2\,(0{,}05)^2(1-0{,}05)^{100-2}$$
$$= \{(100)(99)/(2)(1)\}(0{,}05)^2(0{,}95)^{98}$$
$$= (4.950)\,(0{,}0025)\,(0{,}0066)$$
$$= 0{,}0812$$

UTILIZAREA UNUI CALCULATOR DE DISTRIBUȚIE BINOMIALĂ. Calculatorul de distribuție binomială online (disponibil la http://stattrek.com/tables/binomial.aspx) poate fi utilizat pentru a calcula probabilitățile binomiale individuale și cumulate:

- Introduceți o valoare în fiecare dintre primele trei casete de text.
- Faceți clic pe butonul **Calculați** .
- Calculatorul va calcula probabilitățile binomiale și cumulate .

Probabilitatea de succes la o singură	0,05
Numărul de încercări	100
Numărul de succese (x)	2
Probabilitate binomială: P(X = 2)	0,08118177185776542
Probabilitate cumulativă: P(X < 2)	0,0370812093273546
Probabilitate cumulativă: P(X ≤ 2)	0,11826298118512
Probabilitate cumulativă: P(X > 2)	0,88173701881488
Probabilitate cumulativă: P(X ≥ 2)	0,88173701881488

Probabilitate binomială Pr(*k* = 2) = 0,081181771….

Când ajungeți la capitolul 3 al acestei cărți, puteți intra în mediul de calcul R și puteți utiliza funcția R binom.test() pentru a obține următoarele rezultate: > binom.test(5, 100, p = 0,02, alternativă = „mai mare ", + conf.level=0.05) *# Ieșire:*

TEST BIOMIAL EXACT

date: 5 și 100,
numărul de succese = 5,
numărul de încercări = 100,
p = .05083,
ipoteză alternativă: probabilitatea reală de succes este mai mare decât .02,
CI 5% [0,08919625, 1,00000000],
estimări ale eșantionului: probabilitatea de succes .05.

De asemenea, rezultatele computerului arată că există o probabilitate de 5,083% de a observa cinci sau mai multe complicații severe la 100 de internări. Este compatibil asta cu 2%? Nu există un răspuns definitiv, dar dacă s-ar fi selectat un nivel de semnificație α = 0,05, atunci s-ar concluziona că „da, este compatibil", deoarece $p > .05$, și nu se respinge ipoteza nulă.

Prin abordarea pur și simplu de creștere a dimensiunii eșantionului (adică, numărul de spitalizări), valoarea p devine semnificativă (mai puțin de 0,05), chiar dacă magnitudinea diferenței rămâne aceeași (5% față de 2%) .

Mai târziu, specificând în mod explicit ce diferențe contează, se poate trata acest tip de problemă în mod diferit.

PUTERE

Dacă b = probabilitatea de a nu respinge o ipoteză falsă nulă (adică o eroare de tip II), atunci $(1 - b)$ este probabilitatea de a respinge o ipoteză falsă nulă. Biostatisticienii numesc această probabilitate, $(1 - b)$, **puterea** testului. Este folosit pentru a indica eficacitatea unui test, reflectând faptul că ipoteza nulă este falsificată.

În proiectarea cercetării epidemiologice, este obișnuit să se planifice experimente astfel încât puterea așteptată să fie de 80% sau mai mare, sau puterea de 0,80 sau mai mare, astfel încât investigația să fie eficientă în respingerea unei ipoteze false nule.

CALCULE DE DIMENSIUNEA EȘANTIONULUI PENTRU STUDII ANALITICE (TRIOLA & TRIOLA, 2006)

În testarea ipotezelor cu date eșantion care constau doar din câteva observații, puterea va fi scăzută. Puterea poate fi mărită prin:

- Creșterea dimensiunii eșantionului , alți factori rămânând neschimbați
- Creșterea nivelului de semnificație
- Creșterea abaterii standard
- Utilizarea valorilor extreme pentru parametrii populației

Iată câteva exemple despre calculele dimensiunii eșantionului.

- **Exemplul 1: Într-o** populație dată , trebuie prelevat un eșantion pentru a permite estimarea precisă a mediei eșantionului μ . Determinați dimensiunea eșantionului n necesară.

Soluţie:
Mărimea eșantionului necesar n poate fi estimată din

$$n = (z_{\alpha/2} \, s / E)^2 \qquad (2,4\text{-}10)$$

unde:

$z_{\alpha/2}$ = scorul z critic bazat pe nivelul de încredere dorit, s = abaterea standard a populației, E = marja de eroare dorită.

Observatii:
Pentru a ajunge la rezultatul dat de ecuația (2.4-10) s-au parcurs pașii indicați în următoarele

subsecțiuni.

PASUL 1: ÎNCEPE CU CL T[10]

În teoria probabilității, **CLT** stabilește condițiile în care media unui număr suficient de mare de variabile aleatoare independente, fiecare cu medie și varianță finite, va fi distribuită aproximativ normal.

Un exemplu simplu de CLT este aruncarea unui număr mare de zaruri identice, părtinitoare. Distribuția sumei (sau a mediei) numerelor rulate va fi bine aproximată printr-o distribuție normală. Deoarece cantitățile din lumea reală sunt adesea suma echilibrată a multor evenimente aleatoare neobservate, CLT oferă, de asemenea, o explicație parțială pentru prevalența distribuției normale de probabilitate. De asemenea, justifică aproximarea BIOS-ului cu eșantion mare la distribuția normală în experimentele controlate.

CLT are o serie de variante. În forma sa comună, variabilele aleatoare trebuie să fie distribuite identic. În variante, convergența mediei către distribuția normală are loc și pentru distribuțiile neidentice, având în vedere că respectă condițiile.

CLT poate fi exprimat în diferite forme, în funcție de situația aplicabilă . Două forme comune sunt CLT clasic și CLT Lyapunov.

CLT CLASIC. Fie $X_1, X_2, X_3, \ldots, X_n$ un eșantion aleator de mărimea n - adică o secvență de variabile aleatoare independente și distribuite identic cu valori așteptate m și variațe s^2. Să presupunem că cineva este interesat de comportamentul mediei eșantionului a acestor variabile aleatoare: $S_n = (X_1 + X_2 + X_3, \ldots, + X_n)/ n$. În acest caz, CLT spune că, pe măsură ce n devine mai mare, *distribuția* lui S_n se apropie de normală cu media μ și varianța s^2/n. Adevărata putere a teoremei este recunoașterea ei că S_n se apropie de normalitate indiferent de formele distribuțiilor individuale X_i s.

LIAPUNOV CLT.[11] Fie X_n , $n \in N$ = mulțimea numerelor naturale {1, 2, 3, …}, să fie o succesiune de variabile aleatoare independente. Să presupunem că fiecare X_n are o valoare așteptată finită

$$E[X_n] = m_n \qquad (2,4\text{-}11)$$

[10] CLT (preluat de la http://en.wikipedia.org/wiki/Central_limit_theorem).

[11] CLT lui Lyapunov (preluat de la http://www.enotes.com/topic/Lyapunov's_central_limit teorema).

and finite variance

$$\text{Var}\,[X_n] = s_{n^2} \tag{2.4-12}$$

Also, let

$$S^{N\equiv} \sum_{n=1} {}^{2N_2} s^n \tag{2.4-13}$$

Dacă pentru *unele* $\delta > 0$, valorile așteptate $E[|\,Xk\,|^{2+\delta}]$ sunt finite pentru fiecare $k \in N$ și **starea lui Lyapunov**

$$\lim_{n \to \infty} \frac{1}{S_{ni}^{2+d}} \sum_{i=1}^{n} E^{\lceil} [\,X_i - m_i^{2+d}\,]] = 0 \tag{2,4-14}$$

este mulțumit, atunci CLT-ul ține. Adică variabila aleatoare

$$Z_N := \frac{\sum_{n=1}^{N} (X_n - m_n)}{S_N} \tag{2.4-15}$$

converge în distribuție la o variabilă aleatorie normală standard ca $N \to \infty$ [adică, $Z_N \to N\,(0, 1)$].

Un caz special util al stării lui Lyapunov este atunci când $d = 1$.

Notă : O dovadă viguroasă a CLT poate fi găsită în multe texte clasice despre teoria statisticii matematice. Se crede că cea mai veche descoperire a CLT datează de la omul de știință Laplace în 1776.

PASUL 2: ESTIMAREA O PROPORTȚIE (EȘANȚĂ MARE)[12]

Această metodă urmărește construirea unui CI pentru o proporție de eșantion p . Abordarea descrisă aici este valabilă în următoarele condiții:

- Metoda de eșantionare este eșantionarea simplă aleatorie.
- Eșantionul include cel puțin 5 până la 10 reușite și 5 până la 10 eșecuri .

VARIABILITATEA PROPORȚIEI DE EȘANTION. Pentru a construi un CI pentru o proporție de eșantion , trebuie să cunoaștem variabilitatea proporției de eșantion. Acest lucru permite să se calculeze *abaterea standard și/sau eroarea standard a distribuției de eșantionare.*

Să presupunem că din populație pot fi selectate k eșantioane posibile de mărimea n . Abaterea standard a distribuției de eșantionare este abaterea „medie" dintre k proporțiile eșantionului și proporția reală a populației, P . Abaterea standard a proporției eșantionului s_p este

$$s_p = [\sqrt{\{\,P\,(1 - P)/\,n\,\}}] \times [\sqrt{\{(\,N - n\,)/(\,N - 1)\}}] \tag{2.4-16}$$

unde P este proporția populației, n este dimensiunea eșantionului și N este dimensiunea populației. Când dimensiunea populației este mult mai mare (de cel puțin 10 ori mai mare) decât dimensiunea eșantionului:

$$[\sqrt{\{(\,N - n\,)/(\,N - 1)\}}] \approx 1$$

abaterea standard poate fi aproximată prin

$$s_p = \sqrt{\{\,P\,(1 - P\,)/n\}} \tag{2.4-17}$$

[12] Abaterea standard a proporțiilor eșantionului (preluat de la http://stattrek.com/lesson4/proportion .aspx).

Când proporția reală a populației *P nu* este cunoscută, abaterea standard a distribuției de eșantionare nu poate fi calculată. În aceste circumstanțe, utilizați *eroarea standard* . Eroarea standard (SE) oferă o estimare imparțială a abaterii standard . Poate fi calculat din ecuația 2.4-18 :

$$SE_p = [\sqrt{\{ p\,(1 - p\,)/\,n\,\}}] \times [\sqrt{\{(\,N - n\,)/(\,N - 1)\}}] \qquad (2.4\text{-}18)$$

unde *p* este proporția eșantionului, *n* este dimensiunea eșantionului și *N* este dimensiunea populației. Când dimensiunea populației este de cel puțin 10 ori mai mare decât dimensiunea eșantionului, din nou
$[\sqrt{\{(\,N - n\,)/(\,N - 1)\}}] \approx 1$, iar eroarea standard poate fi aproximată cu

$$SE_p = \sqrt{\{ p\,(1 - p\,)/\,n\,\}} = \sqrt{(\,pq\,/\,n\,)} \qquad (2.4\text{-}19)$$

unde $q = 1 - p$.

PASUL 3: CONCEPTUL DE VALOARE CRITICĂ (TRIOLA & TRIOLA, 2006)

Luați în considerare utilizarea unui scor *z standard* care poate fi utilizat pentru a distinge între eșantionul de BIOS care este probabil să apară și cele care sunt mai puțin probabile sau puțin probabile. Acest scor *z* se numește **valoare critică** și se bazează pe următoarele ipoteze:

(a) În anumite condiții, distribuția de eșantionare a proporțiilor de eșantion poate fi aproximată printr-o distribuție normală (ori de câte ori se aplică CLT).
(b) Proporțiile eșantionului au șanse relativ egale , cu probabilitate de α, de a cădea într-una din cozile curbei normale.
(c) Notând fiecare dintre cele două zone de coadă cu a /2, se vede că există o probabilitate totală de $2(\,a\,/2) = a$ că o proporție de eșantion va aparține uneia dintre aceste două zone de coadă.
(d) Ca complemente ale acestora, există, prin urmare, o probabilitate de $(1 - \alpha\,)$ ca o proporție de eșantion să aparțină zonei principale (necoadă).
(e) Scorul *z* care separă regiunea coadă din dreapta este notat cu $z_{\alpha/2}$ și se numește *valoarea sa critică* deoarece marchează linia limită care separă proporțiile eșantionului care este probabil să apară de cele care nu sunt probabil să apară.

Prin urmare, se poate defini valoarea critică după cum urmează: O **valoare critică** este numărul de pe linia de frontieră care separă BIOS-ul eșantion care este probabil să apară de cele care este puțin probabil să apară.

Numărul $z_{\alpha/2}$ este o valoare critică, care este un scor *z* astfel încât să separă o zonă de a /2 în coada dreaptă a distribuției normale standard.

■ **Exemplu:** Valoarea critică $z_{\alpha/2}$ folosind tabelul *z*

Corespunzător oricărui nivel de încredere, valoarea critică *z* poate fi obținută folosind tabelul *z* .

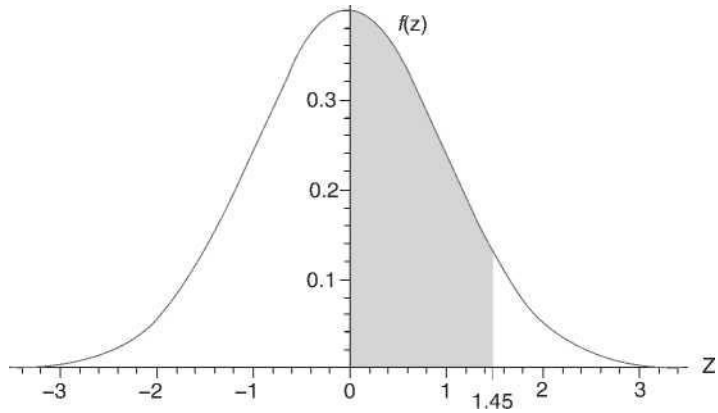


FIGURA 2.5 Curba normală standard și valoarea critică z = 1,45.

În curba normală standard prezentată în Figura 2.5 , media este 0, iar abaterea standard este 1. În diagramă, aria umbrită în gri reprezintă aria care se află în 1,45 abateri standard de la medie. Aria acestei porțiuni umbrite este de 0,4265 (sau 42,65% din suprafața totală de sub curbă).

Pentru a obține această zonă de 0,4265:

1. Citiți în partea stângă a tabelului primele două cifre ale abaterii standard (numărul întreg și primul număr după virgulă, în acest caz 1,4).
2. Citiți peste tabel pentru partea „0,05” (rândul de sus reprezintă al doilea loc zecimal al abaterii standard de interes).

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06
1.4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279

3. Rezultatul este (coloana din stânga) 1,4 + (rândul de sus) 0,05 = 1,45 abateri standard.

Zona reprezentată de 1,45 abateri standard la dreapta mediei este umbrită în verde în curba normală standard din Figura 2.5 .

Tabelul din anexa de la sfârșitul acestui capitol arată cum să găsiți valoarea lui 0,4265 în tabelul *z complet* : Urmați rândul „1.4” de-a lungul și coloana „0.05” în jos până când se întâlnesc la 0.4265.

■ **Exemplu:** Utilizarea tabelului *z* pentru a găsi valoarea critică $z_{\alpha 2}$

În acest exemplu, folosim tabelul *z* pentru a găsi valoarea critică $z_{\alpha/2}$ corespunzătoare unui nivel de încredere de 95%.

Un nivel de încredere de 95% îi corespunde

$$\alpha = 1 - (95\%/100\%) = 1 - 0{,}95 = 0{,}05 \text{ și}$$

$$\alpha/2 = 0{,}05/2 = 0{,}0250$$

corespunde valorii critice a *lui* $z_{\alpha/2}$ — o jumătate de zonă sub curba de distribuție normală standard a

$$0{,}0500 - 0{,}0250 = 0{,}4750.$$

Din tabelul *z* din anexă, aceasta corespunde valorii critice $z_{\alpha/2} = 1{,}96$.

PASUL 4: CONCEPTUL DE MARGE DE EROARE (TRIOLA & TRIOLA, 2006)

Atunci când datele colectate dintr-un eșantion aleatoriu sunt utilizate pentru a estima o proporție a populației p , **marja de eroare** E este diferența maximă probabilă dintre

proporția *p eșantionului observată* și proporția reală a populației *p* , cu probabilitate (1 – *α*). *E* poate fi obținut prin înmulțirea valorii critice și a abaterii standard a proporțiilor eșantionului:

$$E = z\alpha /2 \sqrt{(pq /n)} \quad (2,4\text{-}20)$$

Acest lucru se bazează pe următoarele ipoteze.

Ipoteza 1: Deoarece ambele condiții $np \geq 5$ și $nq \geq 5$ sunt îndeplinite, distribuția de eșantionare a proporțiilor este aproximativ normală. Astfel, se pot folosi rezultate dintr-o altă ipoteză privind distribuția normală (din nou pentru aceste condiții, când se lucrează cu distribuții binomiale): și anume, că variabila aleatoare binomială are o distribuție de probabilitate care poate fi aproximată ca o distribuție normală cu media *m* iar abaterea standard *s* dată de

$$m = np \quad (2,4\text{-}21)$$

și

$$s = \sqrt{(npq)} \quad (2,4\text{-}22)$$

Ambii acești parametri se bazează pe *n* încercări. Astfel, valorile lor pe bază de încercare pot fi obținute, prin împărțirea la *n* , după cum urmează:

Media proporțiilor eșantionului m = (m pentru n încercări)/ $n = np / n = p$ (2,4-23)

Aceasta este *abaterea standard a proporțiilor eșantionului:*

$$s = (s \text{ pentru } n \text{ încercări})/ n = [\sqrt{(npq)}]/ n = \sqrt{(pq / n)} \quad (2.4\text{-}24)$$

PASUL 5: UTILIZAREA REGULI MARJA DE EROARE

Folosind regula marjei de eroare, Ecuația (2.4-20) , cu ipoteza ulterioară ca *p* = Probabilitatea **populației** să fie aproximată prin Probabilitatea **eșantionului** = *p*

$$p \approx p \text{ și } q \approx q$$

astfel încât ecuația (2.4-24) devine

$$\sqrt{(pq)} \approx s \sqrt{(n)} \quad (2,4\text{-}25)$$

iar ecuația (2.4-20) devine

$$E = z_{\alpha/2} \sqrt{(pq /n)} \quad (2,4\text{-}26)$$

$$E = z_{\alpha/2} \sqrt{(pq)} /\sqrt{n}$$

$$= z_{\alpha/2} s \sqrt{(n)} /n$$

$$E = z_{\alpha/2} s /\sqrt{(n)} \quad (2,4\text{-}27)$$

Având în vedere *s* , marja de eroare pentru medie = $E = z_{\alpha/2} (s /\sqrt{ n })$.

PASUL (6) GĂSIREA DIMENSIUNII NECESARE A EȘANȚEI

În cele din urmă, ecuația (2.4-27) arată că dimensiunea eșantionului necesară

$$n = (z_{\alpha/2} s / E)^2 \quad (2,4\text{-}28)$$

depinde de:

- Scorul critic $z_{\alpha/2}$, *care se bazează pe* nivelul de încredere necesar
- Abaterea standard a populației *s*

- Marja de eroare necesară E

Observatii:

1. Ecuația (2.4-28) poate să nu dea un număr întreg pentru n , caz în care, aplicați următoarea **regulă de rotunjire** : *creșteți întotdeauna valoarea mărimii eșantionului n la următorul număr întreg mai mare.*
2. Când utilizați ecuația (2.4-28) , rețineți că aceasta necesită valoarea abaterii standard a populației s , care este de obicei necunoscută. Următoarele metode pot fi utilizate pentru a depăși această problemă:
   (a) Estimați abaterea standard s folosind regula generală a intervalului, după cum urmează: $s \approx$ interval/4.
   (b) Efectuați un studiu pilot prin începerea procesului de eșantionare. Din prima colecție de 30 sau mai multe valori ale eșantionului selectate aleatoriu, calculați abaterea standard a eșantionului s , apoi utilizați s în locul s . Pe măsură ce sunt colectate mai multe date eșantion, valoarea estimată a lui s poate fi îmbunătățită.
3. Utilizați rezultatele altor studii pentru a estima valoarea lui s .
4. Atunci când se calculează dimensiunea eșantionului n , erorile ar trebui să fie conservatoare; a avea n prea mare va da rezultate mai precise decât a-l avea prea mic. De exemplu, dublarea marjei de eroare are ca rezultat scăderea dimensiunii eșantionului la un sfert din valoarea sa inițială. În schimb, înjumătățirea marjei de eroare crește de patru ori valoarea mărimii eșantionului.

> **Exemplu:** Epidemiologia epidemiologilor

Ar trebui să fie evident până acum că epidemiologii sunt profesioniști foarte inteligenți, fiecare cu un IQ ridicat. Testele IQ sunt concepute astfel încât media să fie 100, iar abaterea standard să fie 15. Prin urmare, este de așteptat ca epidemiologii să aibă scoruri IQ cu o medie mai mare de 100 și o abatere standard mai mică de 15, deoarece aceștia reprezintă un grup mai omogen din punct de vedere intelectual decât populatia medie. Pentru a determina IQ-ul mediu pentru populația de epidemiologi, cineva ar dori să știe:

Câți epidemiologi trebuie să fie aleși aleatoriu pentru testarea IQ, dacă se dorește o încredere de 90%, 95% sau 99% că media eșantionului este în 5 puncte IQ față de media populației?

Soluţie:
Folosind ecuația (2.4-28) ,

$$n = (z\alpha_{/2s} / E) 2$$

valorile necesare pentru calcularea mărimii eșantionului n se determină după cum urmează:

$z\alpha_{/2 = 1,645}$. 1,960 și 2,575 pentru 90%, 95% și, respectiv, 99%.

Acestea sunt găsite prin conversia nivelurilor de încredere de 90/95/99% la α = 0,10/0,05/0,01, respectiv, și apoi găsirea valorii critice $z_{\alpha/2}$, așa cum se arată în Pasul 3.

Set E = 2. Acesta este ales astfel încât media eșantionului să fie în 2 unități IQ față de media μ și să facă marja de eroare necesară 2.

s = 15. Acest lucru este specificat în enunțul problemei.

Cu aceste valori, se poate folosi acum ecuația (2.4-28) pentru a calcula n :

$$n = (z_{\alpha/2} s / E)^2 = \{(1,645 \text{ sau } 1,960 \text{ sau } 2,575)(15)/2\}^2$$

$$= 152,21/216,09/372,97$$

care se rotunjesc la 153, 217 sau, respectiv, 373.

Deci, dacă cineva poate găsi un eșantion aleatoriu de 217 epidemiologi și obține IQ-ul fiecăruia, se poate avea un nivel de încredere de 95% că media eșantionului $\underline{x}$ va fi în 2 unități IQ din media reală a populației. Dacă pot fi localizați doar 153 de epidemiologi, mulțumiți-vă cu nivelul de încredere de 90%. Dar dacă pot fi găsiți 373 de epidemiologi, atunci faceți totul pentru nivelul de încredere de 99%. (Vrei să încerci ceva din astea?)

Folosind R

Desigur, se poate folosi mediul R pentru a rula acest calcul după cum urmează: >

```
>   # Testarea IQ a epidemiologilor
>   # 90/95/99 % Calcularea nivelurilor de încredere
>   # Calcularea n = Dimensiunile Eșantionului necesare ale subiecților de caz
>
>   z <- c(1,645, 1,960, 2,575) # z este introdus ca un vector cu 3
>   E <- 2
>   s <- 15
> n <- (z*s/E)**2
>   n # Ieșirea celor 3 dimensiuni de eșantion corespunzătoare:
```

[1] 152.2139 216.0900 372.9727 >

Folosind R , puteți obține aceleași rezultate într-o singură rulare.

- **Exemplu:** Utilizarea unui tabel sau a unui calculator online pentru mărimea eșantionului

Numeroase instrumente de calcul sunt disponibile online. Câteva exemple sunt:

Utilizarea tabelelor: www.itl.nist.gov/div898/handbook/prc/section2/prc222.htm
Folosind un calculator: http://statisticslectures.com/topics/samplesizepopulationmean

Evaluarea unei singure măsuri de apariție (Virasakdi, 2011)

În general, probabilitatea de a întâlni un eveniment epidemiologic semnificativ este foarte mică. Prin urmare, în loc de probabilitate, măsurarea se concentrează pe **densitate** , ceea ce înseamnă incidența sau numărul de apariții într-o perioadă de timp. Timpul este doar o dimensiune; același concept se aplică și pentru densitatea numărului de obiecte mici într-o zonă bidimensională sau spațiu tridimensional.

Când evenimentele sunt independente de altele, apariția este aleatorie. Matematic , se poate dovedi că în această condiție, densitățile în diferite unități de timp variază cu o varianță egală cu densitatea medie. Când probabilitatea de a avea un eveniment este afectată de unii factori, un model poate fi utilizat pentru a explica, analiza și prezice densitatea. Variația între diferitele straturi este explicată de factori. În cadrul fiecărui strat, distribuția este aleatorie.

Număr de Poisson (incidență) și date privind rata

Regresia Poisson se ocupă de variabilele de rezultat care sunt în natură numărătoare (numere întregi sau numere întregi). Covariatele independente sunt similare cu cele întâlnite în regresia liniară și logistică.

În EPDM, regresia Poisson este utilizată pentru analiza datelor grupate de cohortă, analizând densitatea incidenței între persoană-timp contribuită de subiecți cu caracteristici similare de interes. Este unul dintre cele trei modele liniare generalizate (GLM) utilizate în mod obișnuit:

- Regresia Poisson
- Regresia liniară
- Regresia logistică

Există două ipoteze principale pentru regresia Poisson:

În primul rând, riscul este omogen între persoane-timp contribuite de diferiți subiecți care au aceleași caracteristici de interes (de exemplu, gen, grupă de vârstă etc.) și aceeași perioadă.

În al doilea rând, asimptotic (adică, pe măsură ce dimensiunea eșantionului crește), media numărărilor este egală cu varianța.

Distribuția Poisson este o distribuție de probabilitate discretă cu următoarea funcție de densitate:

$$P\left(X = x\right) = \frac{-eu\,X}{x!}l \tag{2,4-29}$$

unde X este variabila aleatoare, x este numărul observat și λ este *numărul așteptat* . Funcția de distribuție este

$$P\left(X \le x\right) = \sum_{k=0}^{x} \frac{K\,{}^{l}\imath}{k!} \tag{2.4-30}$$

Date binomiale de risc și prevalență (Lachin, 2011)

RISC BINOMIAL

În cercetarea epidemiologică, o structură tipică de date constă din n observații independente și distribuite identic (iid) $\{ x_i \}$ dintr-un eșantion de n subiecți de caz ($i = 1, 2, 3, \ldots, n$) extrași la întâmplare dintr-o populație cu probabilitatea p a unei caracteristici de interes, cum ar fi moartea sau deteriorarea sau supraviețuirea sau îmbunătățirea.

Astfel, X este o variabilă binară astfel încât:

$$x_i = \text{I}\{\text{răspuns pozitiv pentru a } i\text{-a observație}\}$$

unde I{.} este funcția indicator: I{.} = 1, dacă este adevărat, sau = 0 dacă nu.

Numărul total de subiecți caz din eșantion cu răspuns pozitiv este $y = \sum_i x_i$, iar proporția simplă cu răspuns pozitiv în eșantion este $p = y / n$. În astfel de cazuri, modelul de probabilitate de distribuție binomială este de obicei modelul de alegere pentru analiza datelor; prin urmare, se introduce **riscul binom .**

DATE DE PREVALENTA

Prevalența unei caracteristici este probabilitatea **P** *ca* caracteristica să apară în populație sau proporția *p* într-un eșantion, cu acea caracteristică prezentă într-o secțiune transversală a populației la un moment dat. De exemplu, prevalența diabetului zaharat de tip 2 la adulți din 1980 a fost estimată la aproximativ 6,8% din populația SUA, pe baza Sondajului Național de Sănătate și Nutriție. Jumătate dintre cei care au îndeplinit criteriile pentru diabet la un test oral de toleranță la glucoză (3,4%) au fost nediagnosticați anterior. În acel studiu, *n* este dimensiunea totală a eșantionului a cărei *y* are caracteristica pozitivă, în acest caz, diabetul de tip 2.

Incidența unui eveniment (caracteristica pozitivă) este probabilitatea *P* în populație, sau proporția *p într-un eșantion, care dobândesc caracteristica pozitivă sau experimentează un eveniment într* - un interval de timp printre cei care nu aveau caracteristica la momentul inițial. . În acest caz, *n* este dimensiunea eșantionului la risc într-un studiu prospectiv de urmărire longitudinală a cui *experimentează* evenimentul într-o perioadă de timp. De exemplu, din sondajul anual National Health Interview, s-a estimat ca incidenta unui nou diagnostic de diabet in randul adultilor din populatia SUA este de 2,42 cazuri noi la 1.000 in populatie pe an.

Astfel de estimări ale prevalenței unei caracteristici sau incidenței unui eveniment de interes sunt, în general, proporții simple bazate pe un eșantion de observații *n iid.*

Evaluarea a două măsuri de apariție — Comparația riscului: raportul de risc și riscul atribuibil

Pentru a compara riscul de îmbolnăvire în diferite grupuri de expunere, pot fi utilizate două metode:

1. **Raportul de risc** (**RR;** numit și **risc relativ**) este *raportul* dintre riscul de îmbolnăvire în rândul celor expuși în comparație cu cel al celor neexpuși. Indică de câte ori ar crește riscul dacă subiectul și-ar schimba statutul de la neexpus la expus. Creșterea este considerată în multipli sau „-fold" și, prin urmare, RR este un model multiplicativ.
2. **Diferența de risc** (**DR)** măsoară *cantitatea* de risc câștigată sau pierdută dacă subiectul trece de la neexpus la expus. Creșterea este absolută și, prin urmare, RD este un model *aditiv* .

RR este un indicator important pentru cauzalitate. Un RR mai mare sau egal cu 10 sugerează puternic o relație cauzală. Cu toate acestea, RD are mai multă sănătate publică implicații decât RR. Un RR ridicat poate să nu fie de importanță pentru sănătatea publică dacă boala este foarte rară. Cu toate acestea, RD măsoară povara directă a sănătății și nevoia de servicii de sănătate.

De exemplu, riscul de a dezvolta cancer pulmonar în rândul celor expuși (de exemplu, fumători) este cu 1 mai mare decât în rândul celor neexpuși (de exemplu, nefumătorii). În această situație , RD își schimbă semnul din pozitiv în negativ. RR

inversează la o valoare mică.

RR crește pe măsură ce crește doza (timpul) de expunere la fumat.

RELAȚIA DOZĂ-RĂSPUNS

Un criteriu de cauzalitate este dovada unei relații doză-răspuns. Dacă o doză mai mare de expunere este asociată liniar cu un nivel mai mare de risc, atunci expunerea este probabil să fie cauza.

Compararea a două estimări de rate: Rate Rate *rr*

Să presupunem că avem un studiu de cohortă care furnizează date binomiale în formatul prezentat în Tabelul 2.8 .

rr raportului ratei corespunzătoare sunt

$$\text{raportul ratei } rr \equiv r_1 / r_0 = (a / PT_1)/(b / PT_0) \qquad (2.4\text{-}31)$$

Compararea a două estimări de risc: raportul de risc *RR și* raportul cotelor de boală (morbiditate) *DOR*

Un studiu de cohortă furnizează date binomiale în formatul prezentat în Tabelul 2.9 .

Relațiile *RR* ale raportului de risc corespunzătoare sunt

$$\text{Raportul de risc } RR \equiv R_1 / R_0 = (a / N_1)/(b / N_0) \qquad (2,4\text{-}32)$$

$$\text{Raportul cotelor de boală } DOR \equiv \frac{R_1/(1-R_1)}{R_0/(1-R_0)} = \frac{(a/N_1)/(1-a/N_1)}{(b/N_0)/(1-b/N_0)} = \frac{a(N_0-b)}{b(N_1-a)} \qquad (2,4\text{-}33)$$

TABELUL 2.8 Notarea tabelului pentru studiul de cohortă, date

	EXPUS	NEEXPUS	TOTAL
Numărul de cazuri noi	*o*	B	M
Persoană-timp în pericol	PT_1	PT_0	*T*

TABELUL 2.9 Studiu de cohortă, date binomiale

	EXPUS	NEEXPUS
Numărul de cazuri noi	*o*	*b*
Persoană aflată în pericol	N_1	N_0

COTE ŞI SAU

Conceptul de cotă este legat de probabilitate: Dacă *p* este probabilitatea, atunci *p* /(1 *p*)

este **cota** . În schimb, probabilitatea *p* ar fi egală cu *cotă/(cotă + 1).*

BOLI (MORBIDITATE) SAU/RATE

În studiile PH, ratele sunt folosite pentru a descrie probabilitatea unui eveniment, pentru a monitoriza starea de sănătate a unei populații și așa mai departe. De obicei, un interval de timp de 1 an este folosit pentru a descrie aceste rate.

O *rată* este frecvența de apariție a unui eveniment semnificativ. Este frecvența relativă a evenimentului înmulțită cu un număr, de obicei o valoare cum ar fi 1.000 sau 100.000. Rata este exprimată ca (x / y) k , unde:

x = numărul de frecvență al numărului de subiecți pentru care a avut loc evenimentul, y = numărul total de persoane expuse riscului de apariție a evenimentului, k = un număr precum 1.000 sau 100.000.

Această definiție poate fi aplicată măsurilor de, de exemplu, mortalitate (decese), morbiditate (boli) și fertilitate (nașteri). În aceste cazuri, k este de obicei luat ca 1.000.

■ **Exemplu de rata de mortalitate:** În Statele Unite, într- un anumit an, au existat 2.416.000 de decese într-o populație de 285.318.000. Care este rata brută de mortalitate?

Rezolvare: Rata brută a mortalității = (2.416.000/285.318.100) × 1.000 = 8,4677. Aceasta este rotunjită la o rată a mortalității de 8,5 decese pentru fiecare 1.000 de persoane din populație .

■ **Exemplu de rata de mortalitate infantilă :** Într -un an recent , în Statele Unite au existat 4.026.000 de născuți vii și 27.500 de decese de sugari sub 1 an. Care este rata mortalității infantile?

Rezolvare: Rata mortalității infantile = (27.500/4.026.000) × 1.000 = 6,8306. Aceasta este rotunjită la o rată a mortalității infantile de 6,8 la 1.000 de copii sub 1 an. Aceasta este substanțial mai mică decât rata mortalității infantile de peste 35 la 1.000 în unele țări.

Compararea a două estimări de șanse de la caz-control: poliomielita Salk Studiu epidemiologic de vaccin

Într-una dintre cele mai renumite investigații epidemiologice (Triola & Triola, 2006) s-a constatat că din 200.745 de copii care au primit vaccinul Salk, 33 au dezvoltat - poliomielita paralizantă. Prin urmare, pentru acest grup de tratament:

$$P(\text{polio}) = 33/200.745 = 0.000.163,9$$

Această măsură unică nu spune întreaga poveste, deoarece această măsură nu conține informații despre rata poliomielitei pentru cei cărora li s-a administrat o injectare placebo. *Riscul de poliomielita la copiii tratați cu vaccinul Salk trebuie comparat cu riscul de poliomielita pentru cei cărora li s-a administrat un placebo.* Acum luați în considerare datele pentru acea investigație rezumate în Tabelul 2.10 .

Pe baza tabelului 2.10 , se pot calcula următoarele probabilități:

Rata poliomielită pentru grupul de tratament = P (poliomielita/vaccin)
= 33/200.745 = 0,000.163,9 = p_1

Rata poliomielită pentru grupul placebo = P (poliomielita/placebo)
= 115/200.229 = 0,000.574,3 = p_2

RR, sau riscul relativ, este

$$p_1 / p_2 = 0,000,163,9/0,000,574,3 = 0,285,4$$

Raportul **de risc reciproc** = 1/0,285,4 = 3,504, ceea ce înseamnă că grupul placebo este de 3,504 ori mai probabil să sufere de poliomielita.

Întrebări de revizuire pentru Secțiunea 2.4

1. (a) Numiți cinci domenii de investigație epidemiologică în care BIOS este utilizat pentru analizele datelor.
   (b) Epidemiologii de cercetare își publică frecvent concluziile în reviste profesionale . Enumerați cinci reviste în care se găsește adesea munca lor.
2. (a) Numiți cinci măsuri epidemiologice de apariție și asociere.
   (b) Cum se obține aceste măsuri în investigațiile PH?
3. (a) În cercetarea epidemiologică, ce sunt măsurătorile cu un singur eșantion? Dați un exemplu pentru fiecare.
   (b) Ce sunt măsurătorile din două eșantioane? Dați exemple.
4. (a) Care sunt următoarele măsuri epidemiologice: rapoarte, proporții și rate? Dați exemple pentru fiecare.
   (b) Care sunt metodele de estimare utilizate în mod obișnuit pentru măsurile epidemiologice ? Dați exemple.

TABELUL 2.10 Investigarea epidemiologică a poliomielitei și a vaccinului Salk

TRATAMENT	poliomielita	FĂRĂ POLIOMIETA	TOTAL
Vaccinul Salk	33	200.712	200.745
Placebo	115	200.114	201.229

5. (a) Care sunt CI în estimarea epidemiologică a dimensiunilor eșantionului? Da un exemplu.
   (b) Utilizând testarea ipotezei nule, cum se pot determina CI corespunzând unui nivel de încredere prescris?
6. (a) Ce este puterea biostatistică?
   (b) Cum este această putere legată de dimensiunea eșantionului unei investigații epidemiologice? Da un exemplu.
7. (a) Care este teorema limită centrală (CLT) în statistica matematică?
   (b) Cum contribuie aceasta la procesul de estimare a mărimii eșantionului necesar într-o investigație epidemiologică? Da un exemplu.
8. (a) Explicați pe scurt conceptele de risc binom și prevalență și dați un exemplu - pentru fiecare.
   (b) Comparați conceptele de RR și diferență de risc și dați un exemplu pentru fiecare.
9. (a) Descrieți utilizarea cotelor și OR, dând un exemplu pentru fiecare.
   (b) În comparația riscurilor, ce sunt un risc atribuibil și RR al acestuia?
10. (a) Ce se înțelege prin rata mortalității, rata mortalității infantile și rata fertilității? Dați un exemplu pentru fiecare.
    (b) Într-o investigație epidemiologică caz-control, care este RR (sau riscul relativ) și

raportul de risc reciproc? Da un exemplu.

EXERCIȚII PENTRU CAPITOLUL 2 (CDC, 2006; BROADBENT, 2011)

Folosind teoria probabilității

1. În fiecare dintre următoarele situații, exprimați gradul de probabilitate indicat ca valoare de probabilitate:
   (a) Prognoza meteo spune că mâine sunt 25% șanse de ninsoare.
   (b) În ghicirea răspunsului pentru cele cinci opțiuni dintr-o întrebare de test cu variante multiple, cineva are mai puțin de 50-50 de șanse de succes.
   (c) Unul nu are mai mult de 1 șansă din 10 milioane de a câștiga la loterie.
   (d) Soarele va rasari maine dimineata.
   (e) La aruncarea a două zaruri (fiecare zar are șase fețe), există 1 șansă din 36 de a obține o pereche de șase.
2. Un cuplu căsătorit are patru copii. Care este probabilitatea fiecăruia dintre următoarele rezultate?
   (a) Toate patru sunt fete.
   (b) Toți patru sunt băieți.
   (c) Trei fete și un băiat.
   (d) O fată și trei băieți.
   (e) Două fete și doi băieți.

Simptomele bolii în studiile clinice de medicamente

3. Un grup de 937 de subiecți de caz a primit un medicament de probă într-un tratament clinic al obezității ; 17 dintre subiecți au prezentat febră mare, dar restul nu.

(a) Care este probabilitatea ca un subiect care ia medicamentul de testat să aibă simptom de febră?
(b) Ar fi considerat normal ca o persoană care ia același medicament să aibă - febră? De ce sau de ce nu?

Riscuri și șanse în epidemiologie

4. Utilizați datele rezumate în Tabelul 2.10 (Investigarea Epidemiologică a Poliomielitisului și a Vaccinului Salk) pentru a calcula:
   (a) reducerea absolută a riscului care poate fi utilizată pentru a evalua eficacitatea vaccinului Salk
   (b) riscul relativ
   Explicați în cuvinte ce se înțelege prin fiecare rezultat calculat.

Studiu epidemiologic caz-control

5. Într-un studiu caz-control al eficacității căștilor de protecție pentru biciclete pentru prevenirea leziunilor faciale, datele sunt rezumate în următorul tabel:

	CASCA PURTATA	CASCA NU SE PURTĂ
Leziuni faciale primite	30	182
Toate leziunile non-faciale	83	236

(a) Calculați valoarea reducerii riscului absolut pentru leziunile faciale în cele două grupuri.
(b) Pentru cei care nu poartă căști, care sunt șansele pentru leziuni faciale?
(c) Care este OR pentru leziunile faciale în grupul care nu a purtat căști în comparație cu grupul care a purtat căști? Ce înseamnă acest raport?
(d) Purtarea căștii scade riscul de răni faciale? De unde ştiţi?
(e) Care este raportul de risc al rănilor faciale pentru cei care poartă căști? Care este raportul de risc pentru cei care nu poartă căști? Este rezonabil ca purtarea căștii să fie o cerință legală? De ce sau de ce nu?

Mortality, Morbidity, and Fertility Rates

7. In a certain year in the United States, the following epidemiologic BIOS were taken:
   Population: 285,318,000
   Live births: 4,026,000
   Women aged 15–44: 61,811,000
   Deaths from HIV infection: 17,402
   Deaths: 2,416,000
   Infant deaths: 27,500
   HIV-infected persons: 900,000
   Motor vehicle deaths: 43,900

(a) Care este rata mortalității infantile?
(b) Care este rata natalității?
(c) Care este rata mortalității HIV pentru persoanele infectate cu HIV?
(d) Care este rata generală de fertilitate?
(e) Folosind k = 100.000, calculați rata de incidență a decesului autovehiculului.

Ratele de incidență în analiza de supraviețuire a cohortelor de caz

8. O echipă de anchetatori epidemiologici a înrolat 4.200 de femei într-un studiu și le-a urmărit anual timp de 4 ani pentru a determina rata de incidență a bolilor de inimă. ■ După 1 an, niciunul nu a avut un nou diagnostic de boală de inimă, dar 200 au fost pierdute pentru urmărire.
   - ■ După 2 ani, 1 a avut un nou diagnostic de boală de inimă, iar alți 198 au fost pierduți de urmărire.
   - ■ După 3 ani, alți 7 au avut noi diagnostice de boli de inimă, iar 1.586 au fost pierdute pentru urmărire.
   - ■ După 4 ani, alți 8 au avut noi diagnostice de boală de inimă și încă 784 au fost pierdute pentru urmărire.

   Calculați rata de incidență a bolilor de inimă în rândul acestei cohorte. Să presupunem că persoanele cu diagnostice noi de boli de inimă și cele pierdute în urma urmăririi au fost libere de boală timp de jumătate din an și astfel contribuie cu jumătate de an la numitor.
9. Un studiu de urmărire a diabetului a inclus 218 femei diabetice și 3823 femei fără diabet. Până la sfârșitul studiului, 72 dintre femeile diabetice și 511 dintre femeile nediabetice au murit. Femeile diabetice au fost observate pentru un total de 1.862 ani-persona; femeile non-diabetice au fost observate pentru un total de 36.653 ani-persona. Calculați ratele de incidență a decesului la femeile diabetice și nediabetice.

Prevalența

10. Într-un studiu efectuat pe 1.150 de femei care au născut, un total de 468 au raportat că au luat o doză prescrisă de vitamine de cel puțin 4 ori pe săptămână în timpul lunii înainte de a rămâne însărcinate. Calculați prevalența consumului frecvent de vitamine în acest grup.

Ratele de mortalitate

11. Tabelul 2.11 furnizează numărul de decese din toate cauzele și din accidente (- răniri neintenționate) pe grupe de vârstă în 2002 în Statele Unite.
    (a) Calculați următoarele rate de mortalitate:
        (i) Rata mortalității specifice leziunilor neintenționate pentru întreaga populație
        (ii) Mortalitatea de toate cauzele în rândul bărbaților
        (iii) Rata mortalității pentru toate cauzele pentru persoanele cu vârsta între 25 și 34 de ani
        (iv) Mortalitatea specifică vătămărilor neintenționate în rândul bărbaților cu vârsta cuprinsă între 25 și 34 de ani
    (b) Sugerați cum să numiți fiecare rată a mortalității.

TABEL 2.11 Mortalitatea de toate cauzele și vătămările neintenționate și populația estimată în funcție de vârstă
Group, pentru ambele genuri și pentru bărbați singuri, Statele Unite, 2002

TOATE RASELE, AMBELE SEXURI				TOATE RASELE, BĂRBAȚI		
GRUPĂ DE VÂRĂ (ANI)	TOATE CAUZELE	RĂNIRI NEintentionate	POP ESTIMAT. (X 1000)	TOATE CAUZELE	RĂNIRI NEINTENȚIONATE	POP ESTIMAT. (X 1000)
0–4	32.892	2.587	19.597	18.523	1.577	10.020
5–14	7.150	2.718	41.037	4.198	1.713	21.013
15–24	33.046	15.412	40.590	24.416	11.438	20.821
25–34	41.355	12.569	39.928	28.736	9.635	20.203
35–44	91.140	16.710	44.917	57.593	12.012	22.367
45–54	172.385	14.675	40.084	107.722	10.492	19.676
55–64	253.342	8.345	26.602	151.363	5.781	12.784
65+	1.811.720	33.641	35.602	806.431	16.535	14.772
Nu este declarat	357	85	0	282	74	0
Total	2.443.387	106.742	288.357	1.199.264	69.257	141.656

Sursa: Web-based Injury Statistics Query and Reporting System (WISQARS) [baza de date online] Atlanta; Centrul Național pentru Prevenirea și Controlul Leziunilor. Disponibil de la: http://www.cdc.gov./ncipc/wisqars.

Estimarea dimensiunilor eșantionului

12. Factor de corecție a populației finite pentru CI:
Pentru o populație **infinită** , eroarea standard a mediei = $s / \sqrt{n}$
Pentru o populație **finită** N, trebuie aplicat factorul de corecție $\sqrt{\{(N - n)/(n - 1)\}}$, ori de câte ori n > 0,05N. Acest factor ar trebui aplicat în ecuația (2.4-27) :

$$E = z_{\alpha/2}\, s / \sqrt{(n)} \qquad (2,4\text{-}27)$$

astfel încât marja de eroare va fi dată de

$$E = \{ z_{\alpha/2}\, s / \sqrt{(n)}\} \sqrt{\{(N - n)/(n - 1)\}} \qquad (2.4\text{-}34)$$

Găsiți IC de 95% pentru media scorurilor IQ luate pentru 1.000 de epidemiologi, dacă 40 dintre aceste scoruri au produs o medie de 120. Să presupunem că s = 15.

13. Factor de corecție a populației finite pentru dimensiunea eșantionului:
În ecuația (2.4-27) , sa presupus că populația este infinită sau foarte mare și că eșantionarea a fost efectuată *cu înlocuire* . Totuși, dacă populația este mică și eșantionarea a fost efectuată ***fără înlocuire*** , ar trebui modificată marja de eroare E pentru a include un factor de corecție a populației finite, astfel încât E să fie dat de ecuația (2.4-34) . Rezolvând această ecuație pentru n , se obține:

$$n = \{ N s^2 (z_{\alpha/2})^2 \}/\{(N - 1) E^2 + s^2 (z_{\alpha/2})^2 \} \qquad (2.4\text{-}35)$$

Repetați exemplul precedent: Epidemiologia epidemiologilor, presupunând că epidemiologii sunt selectați aleatoriu *fără înlocuire* dintr-o populație de *N* = 300 de epidemiologi reputați.

REFERINȚE

Aragon, TJ (2011). *Epidemiologie aplicată folosind R* (epir). Berkeley, CA: Universitatea din California, Berkeley, Școala de Sănătate Publică și Departamentul de Sănătate Publică din San Francisco.

Broadbent, A. (2011). Deducerea cauzalității în epidemiologie: mecanisme, cutii negre și contraste . În PM Illari, F. Russo și J. Williamson (eds.), *Causality in the sciences.* Oxford, Marea Britanie: Oxford University Press. Preluat de la http://www.hps.cam.ac.uk/people/broadbent/ epidemiologie_mecanisme_inference.pdf

Ephron, E. (1984). *Apocaliptica: Cancerul și marea minciună — Cum controlează politica de mediu ceea ce știm despre cancer* . New York, NY: Simon & Schuster.

Hill, AB (1965). Mediul și boala: asociere sau cauzalitate. *Proceedings of the Royal Society of Medicine, 58* , 295–300.

Horwitz, RI și Feinstein, AR (1978). Metode analitice alternative pentru studii caz-control ale estrogenilor și cancerului endometrial. *New England Journal of Medicine, 299* , 1089–1094.

Lachin, JM (2011). *Metode biostatistice: Evaluarea riscurilor relative* (Seria Wiley în Probabilitate și Statistică). Hoboken, NJ: John Wiley & Sons.

Universitatea Loma Linda, Școala de Sănătate Publică. (2012). *Programe în EPDM* . Preluat de pe http:// www.llu.edu/public-health/programs/mph-epdm-track2-research.page

Paoli, B., Haggard, L. și Shah, G. (2002). *Intervalele de încredere în sănătatea publică* . Utah, UT: Biroul de Evaluare a Sănătății Publice din Utah. Preluat de la http://health.utah.gov/opha/IBIShelp/ ConfInts.pdf

Rothman, KJ (1976). Cauze. *Jurnalul American de Epidemiologie, 104* , 587–592.

Rothman, KJ (1998). *Epidemiologie modernă.* Boston, MA: Lippincott Williams & Wilkins.

Rothman, KJ (2002). *Epidemiologie: o introducere.* New York, NY: Oxford University Press.

Rothman, KJ și Greenland, S. (2005). Cauzalitate și inferență cauzală în epidemiologie. *Jurnalul American de Sănătate Publică, 95* (Supp. 1), S1, S144–S150.

Steiger, JH (2015). *O introducere de bază în inferența statistică* . Preluat de la http://www.stat-power.net/Content/310/A Basic Introduction to Statistical Inference.pdf

Triola, MM și Triola, MF (2006). *Biostatistica pentru stiintele biologice si ale sanatatii.* Boston, MA: Pearson/Addison Wesley.

Departamentul de Sănătate și Servicii Umane din SUA, Centrele pentru Controlul și Prevenirea Bolilor (CDC), Biroul pentru Forța de Muncă și Dezvoltarea Carierei. (2006). *Principii de epidemiologie în practica sănătății publice: o introducere în epidemiologia aplicată și biostatistică* (Curs de auto-studiu SS1000, ed. a 3-a). Atlanta, GA: Autor.

Universitatea din Sydney, Australia, Programul Școlii de Sănătate Publică. (2012). Preluat de pe http://sydney.edu.au/medicine/public-health/future-student/study-program/coursework- grade/structură-clinic-epidemiologie.php

Virasakdi, C. (2011). *Analiza datelor epidemiologice folosind* R *și Epicalc.* Unitatea de Epidemiologie, Prince of Songkla University, Thailanda. Preluat de la cvirasak@medicine.psu.ac.th

APENDICE

Tabelul *z* pentru o distribuție normală standard

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0.2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3304	0,3365	0,3389
1.0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1.1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1.2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1.3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1.4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1.5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1.6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1.7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1.8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1.9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2.0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2.1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2.2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2.3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2.4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2.5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2.6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964

2.7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2.8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2.9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3.0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3.1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3.2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3.3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3.4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3.5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3.6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3.7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3.8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999

THREE

Utilizarea analizei datelor Programare R

INTRODUCERE

Un anunț pentru un loc de muncă vacant pe internet pentru un biostatistician arată după cum urmează:

Rezumatul postului: Biostatistician I

Salariu : Deschis

Angajator : XYZ Research and Biostatistics

Locație: orașul X, statul Y

Tip : Full Time—Nivel de intrare

Categorie : Biometrie/biostatistică, Analiză/prelucrare a datelor, Organizare și administrare statistică

Studii necesare : diplomă de master

XYZ Research and Biostatistics este un lider național în proiectarea, gestionarea și analizarea studiilor clinice pentru cancer. XYZ colaborează cu anchetatorii clinici pentru a oferi expertiză biostatistică respectată, susținută de sisteme sofisticate de gestionare a datelor bazate pe web. Serviciile XYZ asigură implementarea în timp util și sigură a testelor și analize de date fiabile.

Descrierea postului:

Rezumatul poziției : O oportunitate interesantă este disponibilă pentru un biostatistician de a se alătura unui grup mic, dar în creștere, axat pe studiile clinice privind cancerul și cercetarea translațională aferentă. XYZ, care este situat în centrul orașului XX, este responsabil pentru proiectarea, gestionarea și analiza unei varietăți de studii clinice de cancer de fază I, faza II și faza III, precum și analiza datelor de laborator asociate, inclusiv microarray, SNP și date de proteomică. Candidatul de succes va colabora cu colegii de biostatistică și anchetatori clinici pentru a proiecta, evalua și interpreta studiile clinice.

Atribuții și responsabilități principale : analizează studiile clinice și studiile auxiliare asociate, în colaborare cu colegii statisticieni și alți oameni de știință. Pregătește tabele, cifre și rezumate scrise ale rezultatelor studiului; interpretează rezultatele în colaborare cu alți oameni de știință; și ajută la pregătirea manuscriselor. Oferă consultanță statistică cu personalul colaborator. Îndeplinește alte sarcini legate de post, așa cum i-au fost atribuite.

Cerințe:

Calificări necesare : diplomă de master în statistică, biostatistică sau un domeniu conex. Cunoștințe solide de statistică aplicată. Competență în calcul statistic în SAS.

Responsabilități/Calificări preferate : Experiență în consultanță biostatistică. Experiență în limbajul de programare S-Plus sau R. Experiență în analiza datelor cu dimensiuni mari. Abilitatea de a comunica bine oral și în scris. Abilități excelente interpersonale/de lucru în echipă pentru o colaborare eficientă. Cunoștințe de limba spaniolă un plus.

*În scrisoarea de intenție, descrieți modul în care abilitățile și experiența dvs. se potrivesc cu calificările pentru post.

Pentru a afla mai multe despre XYZ, vizitați www.XYZ.org (Asociația Americană de Statistică, nd).

În mod clar, oricine planifică o carieră în biostatistică ar trebui să fie conștient de cerința deschisă a unui nivel acceptabil de competență profesională în analiza datelor folosind mediul de programare R. Chiar dacă cineva nu este un biostatistician care lucrează în domeniile epidemiologiei, sănătății publice și medicinei preventive, un set de abilități care include programarea R ar fi de ajutor.

3.1 DATE ȘI PRELUCRARE A DATELOR

Datele sunt fapte sau cifre din care se pot trage concluzii. Atunci când datele au fost înregistrate, clasificate și organizate, legate și interpretate într-un cadru astfel încât sensul să apară, ele devin **informații** . Există mai mulți pași implicați în transformarea datelor în informații, iar acești pași sunt cunoscuți sub denumirea de **procesare a datelor** . Această secțiune descrie procesarea datelor și modul în care computerele efectuează acești pași în mod eficient și eficient. Multe dintre aceste activități de procesare pot fi efectuate folosind programarea R sau efectuate într-un mediu R cu ajutorul pachetelor R disponibile în care sunt stocate funcțiile și seturile de date R.

Diagrama simplificată care urmează arată modul în care datele brute sunt transformate în informații (Statistics Canada, 2013) :

Date → Colectare → Procesare → Informații

Prelucrarea datelor are loc odată ce toate datele relevante au fost colectate. Acestea sunt colectate din diverse surse și introduse într-un computer unde pot fi procesate pentru a produce **informații** - rezultatul.

Procesarea datelor include următorii pași, fiecare dintre care va fi discutat în secțiunile următoare:

- Codificarea datelor
- Captarea datelor
- Editare
- Imputare
- Controlul calității
- Producerea de rezultate

Codarea datelor

Înainte ca datele brute să poată fi introduse într-un computer, acestea trebuie mai întâi să fie codificate. Pentru a face acest lucru, răspunsurile la sondaj trebuie să fie etichetate; de obicei se folosesc coduri numerice simple. Etichetarea se poate face de către intervievatorul pe teren sau de către un angajat de birou. Etapa de codificare a datelor este importantă deoarece facilitează introducerea și procesarea datelor. Sondajele folosesc două tipuri de întrebări: *închise* și *deschise* . Răspunsurile la aceste întrebări afectează tipul de codare efectuat. Cu o **întrebare închisă** , este permis doar un număr fix de răspunsuri predeterminate la sondaj. Aceste răspunsuri vor fi deja codificate. Următoarele, extrase dintr-un sondaj privind activitățile sportive, sunt un exemplu de întrebare închisă:

În ce măsură sportul este important pentru a vă oferi următoarele beneficii?
< 1/> Foarte important

< 2/> Oarecum important
< 3/> Nu este important

Când sunt folosite **întrebări deschise , orice răspuns este permis, ceea ce face codificarea ulterioară mai dificilă.** Pentru a codifica o întrebare deschisă, procesorul trebuie să eșantioneze un număr de răspunsuri și apoi să proiecteze o structură de cod care să includă toate răspunsurile posibile.

Următoarea structură de cod este un exemplu de întrebare deschisă:

La ce sporturi participi?
Specificați (28 de caractere) _____

În recensământul din SUA și în aproape toate celelalte anchete, codurile pentru fiecare câmp de întrebare sunt premarcate pe chestionar. Când chestionarul este procesat, codurile sunt introduse direct în baza de date și pregătite pentru captarea datelor. Următorul este un exemplu de codare premarcată:

Ce limbă vorbește această persoană cel mai des acasă?
<18/> engleză
<19/> franceza
<20/> Altele—Specificați _____

SISTEME AUTOMATE DE CODARE

Sunt disponibile programe pentru automatizarea sarcinilor repetitive și de rutină. Unele dintre avantajele unui sistem de codificare automatizat sunt că procesul devine din ce în ce mai rapid, mai consistent și mai economic.

Următorul pas în procesarea datelor este introducerea datelor codificate într-o bază

de date computerizată; aceasta se numește *captura de date* .

Captură de date

Captarea datelor este procesul prin care datele sunt transferate de pe hârtie, cum ar fi chestionare și răspunsuri la sondaj, într-un fișier electronic de pe un computer. Înainte ca această procedură să aibă loc, chestionarele trebuie *pregătite* (pregătite) pentru captarea datelor. În această etapă de procesare, fiecare chestionar este revizuit pentru a se asigura că toate datele minime necesare au fost raportate și că sunt descifrabile. Îngrijirea este de obicei efectuată prin editări automate extinse.

Pentru captarea datelor sunt utilizate mai multe metode:

- **Diagramele de calcul** sunt folosite pentru a înregistra date precum numărul de apariții ale unui anumit eveniment și pentru a dezvolta tabele de distribuție a frecvenței.
- **Tastarea batch** este una dintre cele mai vechi metode de captare a datelor; datele sunt introduse prin tastatura computerului. Acest proces este foarte practic pentru intrarea în volum mare, unde producția rapidă este o cerință. Nu sunt necesare proceduri de editare , dar trebuie să existe un grad ridicat de încredere în programul de editare.
- **Captura interactivă** este adesea denumită *tastare inteligentă* . De obicei, datele capturate sunt editate înainte de a fi introduse. Cu toate acestea, captura interactivă combină capturarea și editarea datelor într-o singură funcție.
- **Cititoarele optice de caractere** sau **scanerele de coduri de bare** sunt capabile să recunoască caracterele alfa sau numerice sau codurile de bare. Acești cititori scanează linii și le traduc în program. Scanerele de coduri de bare sunt destul de comune și adesea folosite în magazine. Ele pot lua forma unui pistol de mână sau a unei baghete, precum și a „trece peste” plăci de sticlă.
- **Înregistrările magnetice** au atât capacități de citire, cât și de scriere. Înregistrarea magnetică poate fi utilizată în zonele în care securitatea datelor este importantă. O aplicație importantă pentru acest tip de captare a datelor este banda magnetică care se găsește pe cardurile de debit și de credit.

O tastatură de computer este unul dintre cele mai cunoscute dispozitive de intrare (sau de introducere a datelor) în uz curent. În trecut, oamenii efectuau introducerea datelor folosind carduri perforate sau bandă de hârtie. Alte exemple de dispozitive moderne de introducere a datelor sunt:

- Cititor optic de marcaj
- Cititor de coduri de bare
- Scaner utilizat în desktop publishing
- Pix ușor
- Trackball
- Mouse

Odată ce datele au fost introduse într-o bază de date computerizată, următorul pas este să vă asigurați că toate răspunsurile sunt corecte; aceasta necesită **editarea datelor** .

Editarea datelor

Datele trebuie editate înainte de a fi prezentate ca informații. Această acțiune asigură că informațiile furnizate sunt exacte, complete și consecvente. Există două niveluri de editare a datelor: *microeditare* și *macroeditare* .

Microeditarea corectează datele la nivel de înregistrare. Acest proces detectează erori în date prin verificări ale înregistrărilor individuale de date. Intenția în acest moment este de a determina consistența datelor și de a corecta înregistrările individuale de date.

Macroeditarea detectează și erorile în date, dar face acest lucru prin analiza datelor agregate (totaluri). Datele sunt comparate cu datele din alte anchete, fișiere administrative sau versiuni anterioare ale acelorași date. Acest proces determină - compatibilitatea datelor.

Imputații

Editarea adaugă puțină valoare la îmbunătățirea generală a rezultatelor efective ale sondajului dacă nu se iau măsuri corective atunci când articolele nu respectă regulile stabilite în timpul procesului de editare. Când toate datele au fost editate folosind regulile aplicate și se constată că un fișier are date lipsă, atunci **imputarea** se face de obicei ca un pas separat.

Non-răspunsul și datele nevalide afectează cu siguranță calitatea rezultatelor sondajului. Imputarea rezolvă problemele răspunsurilor lipsă, invalide sau incomplete identificate în timpul editării, precum și orice erori de editare care ar fi putut apărea. În această etapă, toate datele sunt verificate pentru erori, deoarece respondenții nu sunt singurii capabili să facă greșeli; erori pot apărea și în timpul codării și editării.

Alte tipuri de metode de imputare includ următoarele:

- **Hot deck** folosește alte înregistrări ca „donatori” pentru a răspunde la întrebarea (sau la un set de întrebări) care necesită imputare.
- **Cold Deck** utilizează un set fix de valori care acoperă toate elementele de date. Aceste valori pot fi construite folosind date istorice, expertiză în materie și așa mai departe.
- **Înlocuirea** se bazează pe disponibilitatea datelor comparabile . Datele imputate pot fi extrase din fișa respondentului dintr-un ciclu anterior al anchetei sau pot fi preluate dintr-un fișier sursă alternativ (de exemplu, fișiere administrative sau alte fișiere de anchetă pentru același respondent).
- **Estimatorul** folosește informații din alte întrebări sau din alte răspunsuri (din ciclul curent sau dintr-un ciclu anterior), iar prin operații matematice, derivă o valoare plauzibilă pentru câmpul lipsă sau incorect.

Datele donatorilor pot fi găsite și printr-o metodă numită **imputarea celui mai apropiat vecin** . În acest caz, trebuie dezvoltat un fel de criteriu pentru a determina, în conformitate cu caracteristicile predeterminate, care unitate de răspuns este „cel mai asemănătoare” cu unitatea cu valoarea lipsă. Unitatea cea mai apropiată de valoarea lipsă este apoi utilizată ca donor.

Metodele de imputare pot fi efectuate automat, manual sau în combinație.

Calitatea datelor

Calitatea este un element esențial la toate nivelurile de prelucrare. Pentru a asigura calitatea unui produs sau serviciu în activitățile de dezvoltare a sondajelor, se folosesc atât *metode de asigurare a calității,* cât și metode *de control al calității* .

ASIGURAREA CALITĂȚII

Asigurarea calității se referă la toate activitățile planificate necesare pentru a oferi încredere că un produs sau serviciu își va satisface scopul și nevoile utilizatorilor. În contextul activităților de desfășurare a anchetei, activitățile și verificările de asigurare a calității pot avea loc în oricare dintre etapele majore ale dezvoltării anchetei: planificare, proiectare, implementare , procesare, evaluare și diseminare.

Asigurarea calității:

- Anticipează problemele înainte ca acestea să apară
- Utilizează toate informațiile disponibile pentru a genera îmbunătățiri
- Nu este legat de un anumit standard de calitate
- Se aplică mai ales în faza de planificare
- Este atotcuprinzător în activitățile sale

CONTROLUL CALITATII

Controlul calității este o procedură de reglementare prin care:

- Măsoară calitatea
- Compară calitatea cu standardele prestabilite
- Acționează asupra diferențelor

Câteva exemple în acest sens includ controlul calității operațiunii de codificare, calitatea interviului prin sondaj și calitatea captării datelor.

Controlul calitatii:

- Răspunde la problemele observate
- Utilizează măsurătorile continue pentru a lua decizii cu privire la procese sau produse
- standard de calitate prespecificat pentru comparație
- Se aplică mai ales în faza de procesare
- procedură stabilită care este un subset al asigurării calității

MANAGEMENTUL CALITĂŢII ÎN AGENŢIILE STATISTICE

Calitatea datelor trebuie definită și asigurată în contextul „apte pentru utilizare”; această adecvare va depinde de funcția preconizată a datelor și de caracteristicile fundamentale ale calității. Depinde și de așteptările utilizatorilor și de ceea ce ei consideră că sunt informații utile.

Nu există o definiție standard între agențiile de statistică pentru termenul *de statistică oficială* . Există o gamă general acceptată, dar în evoluție, de probleme de calitate sub conceptul de aptitudine pentru utilizare. Aceste elemente de calitate trebuie luate în considerare și echilibrate în proiectarea și implementarea programului statistic al unei agenții.

Următoarea este o listă a elementelor de *calitate* :

- Relevanţă

- Precizie
- Promptitudine
- Accesibilitate
- Interpretabilitate
- Coerenţă

Aceste elemente de calitate tind să se suprapună. Așa cum nu există o măsură unică a acurateței, nu există un model statistic eficient pentru a reuni toate aceste caracteristici ale calității într-un singur indicator. De asemenea, cu excepția cazurilor simple sau unidimensionale (1D), nu există un model statistic general pentru a determina dacă un anumit set de elemente de calitate oferă o calitate generală mai mare decât altul.

Producerea de Rezultate

După editare, datele pot fi procesate în continuare pentru a produce rezultatul dorit. Software- ul de calculator folosit pentru procesarea datelor va depinde de forma de ieșire necesară. În mod obișnuit, sunt utilizate aplicații software pentru procesarea de text, publicare desktop, grafică (inclusiv grafică și desen), programare, baze de date și foi de calcul. Următoarele sunt câteva exemple de moduri prin care software-ul poate produce date:

- **Foile de calcul** sunt programe care adaugă automat coloane și rânduri de cifre, calculează medii și efectuează analize statistice.
- **Bazele de date** sunt dulapuri electronice . Ele stochează în mod sistematic datele pentru un acces ușor și pot produce rezumate, agregate sau rapoarte. Bazele de date relaționale facilitează vizualizarea și compararea subseturilor selectate de date.
- Pot fi dezvoltate **programe specializate pentru a edita, curăța,** imputa și procesa rezultatul final.

Întrebări de revizuire pentru Secțiunea 3.1

1. Ce limbaje de calcul statistice de bază au fost notate ca fiind importante în exemplul de descriere a postului pentru un biostatistician la nivel de intrare? De ce?
2. Într-o școală tipică de sănătate publică sau o școală de medicină, ar trebui ca curriculumul de bază pentru un program tipic de master în sănătate publică în epidemiologie și sănătate publică sau pentru o diplomă în medicină preventivă să includă dezvoltarea competenței în utilizarea R programare pentru biostatistică? De ce sau de ce nu?
3. (a) Contrastați conceptele de date și informații.
   (b) Cum sunt convertite datele în informații?
4. În pașii care convertesc datele în informații, cum sunt aplicate statisticile și calculele diferitelor etape de prelucrare a datelor?
5. (a) Descrieți și delimitați asigurarea calității și controlul calității în prelucrarea datelor computerizate.
   (b) În ce mod apar statisticile în aceste faze ale prelucrării datelor?

3.2 ÎNCEPUT R

R este un mediu software integrat, cu sursă deschisă, disponibil gratuit, pentru manipularea datelor , calcul, analiză și afișare grafică. Mediul R este format din:

- O facilitate de manipulare și stocare a datelor
- Operatori pentru calcule pe matrice și matrice
- O colecție de instrumente pentru analiza datelor
- Capacități grafice pentru analiză și afișare
- Un limbaj de programare eficient și în continuă dezvoltare, asemănător algebrei, care constă din bucle, condiționale, funcții definite de utilizator și capacități de intrare și ieșire

Termenul *de mediu* este folosit pentru a arăta că R este într-adevăr un sistem planificat și coerent (Aragon, 2011; Venables et al., 2004).

R și Biostatistică

R a fost scris inițial de **R** obert Gentleman și **Ross** Ihak a**** de la Departamentul de Statistică al Universității din Auckland, Noua Zeelandă, în 1997. De atunci, un grup central de dezvoltare R de aproximativ 20 de persoane are acces în scriere la codul sursă R.

Mediul original R , a evoluat din limbajele S/S-PLUS , *nu a fost îndreptat* în primul rând către statistică și biostatistică. Cu toate acestea, de la dezvoltarea sa în anii 1990, a devenit instrumentul preferat al multora care lucrează în domeniile tehnicilor statistice clasice și moderne, inclusiv al multora care îl aplică în biostatistică cu privire la epidemiologie, sănătate publică și medicina preventivă (Ara gon, 2011 ; 2006; & Domonici, 2008; Teetor, 2011; Venable şi colab., 2004; Virasakdi, 2011; Verzani, 2005) . Aceste din urmă aplicații sunt *rațiunea de a fi* pentru aceasta carte .

În momentul scrierii acestui articol, cea mai recentă versiune de R este R-2.14.1, lansată oficial pe 22 decembrie 2011. Sursa principală a pachetelor R este Comprehensive R Archive Network (CRAN), la http://cran.r -project.org.

Pachetele R pot fi găsite și în numeroase publicații, cum ar fi *Journal of Statistical Software.* Al 45-lea volum al jurnalului este disponibil la www.jstatsoft.org/v45.

Acum vom începe cu mediul versiunii R-2.9.1, descarcând-o de pe Internet și aruncând o primă privire asupra mediului de calcul R. Amintiți-vă din capitolul 1 că mediul R a fost obținut după cum urmează:

Accesați Internetul pe site-ul CRAN (http://cran.r-project.org).

***Pentru a instala* R *:* R-2.9.1-win32.exe**
http://www.r-project.org/
= > descărcați R
= > Selectați: SUA
http://cran.cnr.Berkeley.edu < http://cran.cnr.berkeley.edu/ >
Universitatea din California, Berkeley, CA
= > http://cran.cnr.berkeley.edu/

**** Prin urmare, programul se numește „ R ”.

= > Windows (95 și mai târziu)
= > baza
= > R-2.9.1-win32.exe
DUPĂ finalizarea descărcării:
= > Faceți dublu clic pe: R-2.9.1-win32.exe
(pe desktop) pentru a dezarhiva și a instala R
= > O pictogramă (Script R 2.9.1) va apărea pe desktopul computerului dvs., așa cum se arată în Figura 3.1 .



FIGURA 3.1 Pictograma R de pe desktopul computerului.

În această carte, următoarea schemă este utilizată pentru toate declarațiile în timpul activităților de calcul din mediul R pentru a clarifica diferitele intrări și ieșiri din procesul de calcul:

1. Textul obișnuit al cărții
2. Introducere linie în codul R
3. **Ieșire de linie în cod** R
4. *Instrucțiuni de comentariu de linie în codul* R

Notă: Semnul **#** este caracterul de comentariu: tot textul dintr-o linie care urmează acestui semn este tratat ca un comentariu de către programul R și nu va fi luată nicio acțiune de calcul cu privire la o astfel de declarație. Adică, activitățile de calcul vor ignora comentariile și vor continua ca și cum declarațiile de comentarii nu ar exista. Aceste declarații de comentariu ajută programatorul și utilizatorul oferind unele clarificări cu privire la scopurile restului mediului R. Cu toate acestea, calculele vor continua chiar dacă aceste declarații de comentariu sunt eliminate.

Pentru a utiliza R cu un sistem de operare Microsoft Windows, faceți dublu clic pe pictograma R 2.9.1.

După ce selectați și faceți clic pe R , se deschide fereastra R , cu următoarea declarație :

R **2.9.1 (26.06.2009)**
Drepturi de autor (C) 2009 The R Foundation for Statistical Computing ISBN 3-900051 07-0
R **este un software gratuit și nu vine cu ABSOLUT NU GARANȚIE.**
Sunteți binevenit să-l redistribuiți în anumite condiții.
Tastați „license()" sau „license()" pentru detalii despre distribuție.
R **este un proiect de colaborare cu mulți colaboratori.**
Tastați „contributors()" pentru mai multe informații și „citation()" despre cum să

citați pachetele R **sau** R în publicații.
Tastați „demo()" pentru unele demonstrații, „help()" pentru ajutor on-line sau „help.start()" pentru o interfață de browser HTML pentru a ajuta.

Tastați „q()" pentru a părăsi R .
[Spațiul de lucru salvat anterior a fost restaurat]

```
>  Acesta este mediul de calcul R.
>  # Calculele pot începe acum!
>
>  În primul rând, să folosim R ca calculator și să încercăm o aritmetică simplă
>  # operație, să spunem: 1 + 1
> 1+1
> 1] 2 # Aceasta este ieșirea!
>  # WOW! Chiar funcționează !
>  # [1] din fața rezultatului de ieșire face parte din modul lui R de a tipări numerele
>  # și vectori. Deși nu este atât de util aici, devine așa atunci când
>  # rezultatul de ieșire este un vector mai lung: vezi exemplul în Secțiunea 3.5.
>
```

Din acest moment, această carte este citită cel mai bine cu mediul R la îndemână. Va fi cea mai eficientă experiență de învățare dacă exersați fiecare comandă R pe măsură ce parcurgeți materialele textuale.

O primă sesiune folosind R

Această subsecțiune introduce câteva caracteristici importante și practice ale mediului R (Venables et al., 2004) .

Conectați-vă și începeți o sesiune R în sistemul Windows al computerului: >

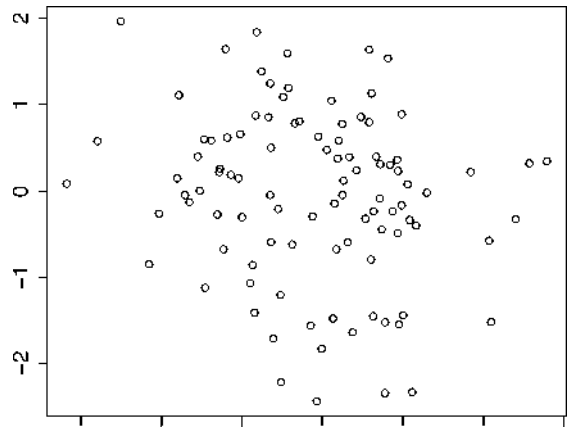
```
>  # Acesta este mediul R.
>  help.start() # Aceasta afișează o pagină care listează diferitele manuale de ajutor online
>                # și materiale disponibile, cum ar fi
>                # Manuale de analiză a datelor statistice ,
>                # „ O introducere în R " ( Venables și colab., 2004)
pornirea serverului de ajutor httpd... gata
Dacă nu se întâmplă nimic, ar trebui să deschideți
„ http://127.0.0.1:28103/doc/html/index.html" dvs.
```

În acest moment, explorați interfața Hypertext Markup Language (HTML) pentru ajutor online chiar de pe desktop, folosind indicatorul mouse-ului pentru a nota diferitele caracteristici ale acestei facilități disponibile în mediul R.

Acum se poate accesa fiecare dintre aceste pachete de programe R și le poate folosi pentru aplicații suplimentare, după cum este necesar.

Revenind la mediul R :

```
>  x <- rnorm(100)
>  # Generarea unui x pseudo-aleatoriu de 100 de vectori
> y <- rnorm(x)
>  # Generarea unui alt pseudo-aleatoriu de 100 de vectori y
>  grafic (x, y)
```



> *# Trasarea* x *vs.* y *în plan, rezultând o fereastră grafică* (Figura 3.2).

Observație: Pentru referință, anexa de la sfârșitul acestui capitol conține documentația CRAN a funcției R plot() , disponibilă pentru ieșire grafică, care poate fi găsită de segmentul de cod R :

> ?complot

CRAN are documentație pentru multe funcții și pachete R. Revenind din nou la spațiul de lucru R , introduceți:

>
> ls() *# (Acesta este un „L" mic urmat de „s"; este comanda „listă".)*
> *# (NU 1 = „UNUL" urmat de „s")*
> *# Această comandă va lista toate obiectele* R *acum în spațiul de lucru* R.
> *# Ieșire:*

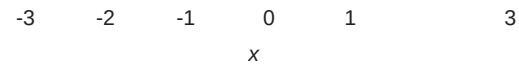


FIGURE 3.2 Graphical output for plot (x, *y*).

> 1] "E" "n" "s" "x" "y" "z"

Reveniți la spațiul de lucru R și introduceți:

>
> rm (x, y) *# Eliminarea tuturor* x *și a tuturor* y *din spațiul de lucru* R
> x *# Apel pentru* x
Eroare: obiectul „x" nu a fost găsit *# Desigur , x - urile tocmai au fost eliminate!*
> y *# Apel pentru* y
Eroare: obiectul „y" nu a fost găsit *# Deoarece* și au *fost eliminate !*
>
> x <- 1:10 *# Fie* x = *(1,2,3,4,5, 6, 7,8,9,10)*
> x *# Ieșirea* x *(doar se verifică!)*
> 1] 1 2 3 4 5 6 7 8 9 10
> w <- 1 + sqrt(x)/2 *# w este un vector de ponderare al abaterilor standard*
> dummy <- data.frame (x = x, y = x + rnorm(x)*w)

> *# Realizarea unui cadru de date din două coloane,* x *și* y *, pentru inspecție*
> dummy *# Ieșirea cadrului de date dummy*

	x	y
1	1	1.311612
2	2	4.392003
3	3	3,669256
4	4	3,345255
5	5	7,371759
6	6	-0,190287
7	7	10,835873
8	8	4,936543
9	9	7,901261
10	10	10,712029

> fm <- lm(y~x, data=dummy) *# Efectuarea unei regresii liniare simple* > summary(fm) *# Ajustarea unei regresii liniare simple a lui* y *pe* x *,* > *# apoi inspectarea analizei și scoaterea în evidență:*

Apel:
lm(formula = y ~ x, date = dummy)

Reziduuri:

Min	1Q	Median	3Q	Max
–6,0140	–0,8133	–0,0385	1,7291	4,2218

Coeficienți:

	Estima	Std. Eroare	valoarea	Pr(>|t|)
(Intercepta)	1,0814	2,0604	0,525	0,6139
x	0,7904	0,3321	2.380	0,0445 *

Signif. coduri: 0 '*' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 " 1**

Eroare standard reziduală: 3.016 pe 8 grade de libertate
R-pătrat multiplu: 0,4146, R-pătrat ajustat: 0,3414
F-statistică: 5,665 pe 1 și 8 DF, valoare p: 0,04453

> fm1 <- lm(y~x, date=dummy, weight=1/w^2)
> summary(fm1) *# Cunoașterea abaterii standard, apoi efectuarea unei ponderi*
> *# regresie și ieșire:*

Apel:
lm(formula=y ~ x, date=dummy, greutate=1/w^2)

Reziduuri:

Min	1Q	Median	3Q	Max
–2,69867	–0,46190	–0,00072	0,90031	1,83202

Coeficienți:

	Estima	Std. Eroare	valoarea	Pr(>|t|)
(Intercepta)	1,0814	2,0604	0,525	0,6139
x	0,7904	0,3321	2.380	0,0445 *

Signif. coduri: 0 '*' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 " 1**

Eroare standard reziduală: 1,356 pe 8 grade de libertate
R-pătrat multiplu: 0,4424, R-pătrat ajustat: 0,3728

F-statistică: 6,348 pe 1 și 8 DF, valoare p: 0,03583

```
> attach(dummy) # Mascarea coloanelor din cadrul de date ca variabile
```

Următoarele obiecte sunt mascate _de_ „.GlobalEnv": x

> lrf <- lowess(x, y) # *O funcție de regresie locală neparametrică lrf* > plot (x, y) # *Realizarea unui grafic punctual standard , rezultând* Figura 3.3 .



FIGURA 3.3 Un grafic punctual standard.

> lines(x, lrf$y) # *Adăugarea liniei locale de regresie*
> # *ieșire* Figura 3.4 .

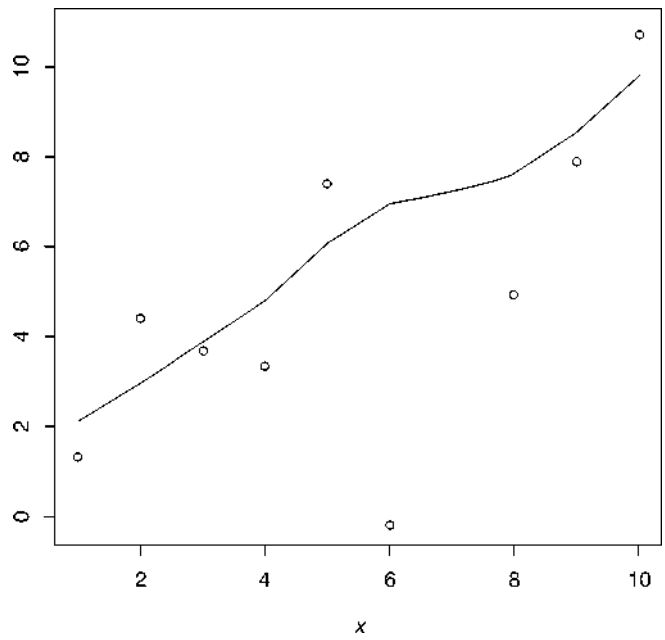


FIGURE 3.4 Adding in the local regression line.

```
> abline(0, 1, lty = 3) # adăugând linia de regresie adevărată:
>                       # (Interceptare = 0, Pantă = 1)
>                       # ieşire Figura 3.5 .
```

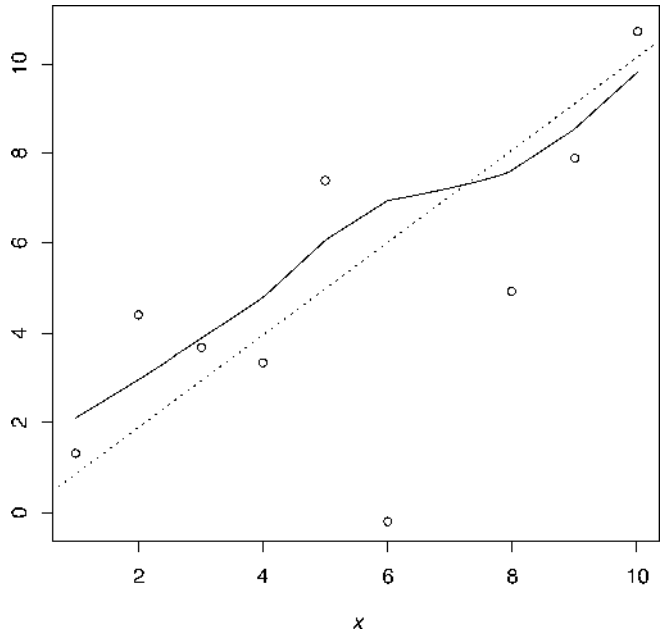


FIGURA 3.5 Adunarea dreptei de regresie adevărată (intersecție = 0, panta = 1).

```
> abline(coef(fm)) # adăugând în linia de regresie neponderată
>                  # ieşire Figura 3.6 .
```

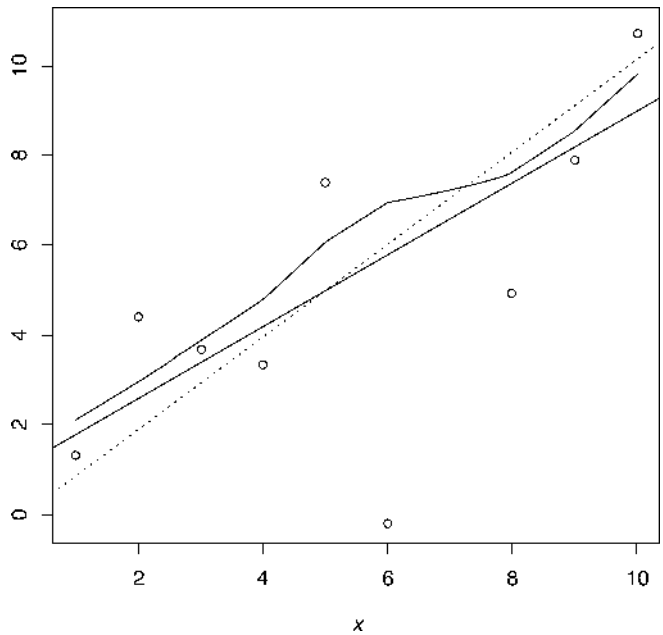


FIGURE 3.6 Adding in the unweighted regression line.

> abline(coef(fm1), col="red") *# adăugând în linia de regresie ponderată* *# ieşire* Figura 3.7 .

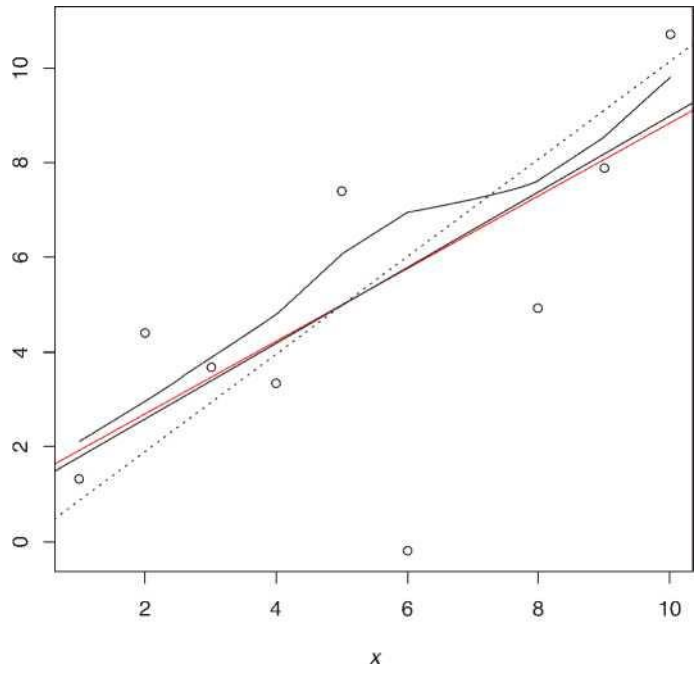


FIGURA 3.7 Adăugarea liniei de regresie ponderată.

> detach() # Se elimină datele cadru din calea de căutare
> plot(instalat(fm), rezid(fm), *# Efectuarea unui diagramă de*
+ xlab="Valori ajustate", *diagnostic standard # pentru a*
+ ylab="reziduale", *verifica heteroscedasticitatea*
+ main="Reziduale vs. montate") *††††, # verificarea variaţiei*
> *diferite # ieşirea în* Figura 3.8 .
rm(fm, fm1, lrf, x, dummy)
fm
Eroare: obiectul „fm" nu a fost găsit
fm1
Eroare: obiectul „fm1" nu a fost găsit
lrf
Eroare: obiectul „lrf" nu a fost găsit
> x
Eroare: obiectul „x" nu a fost găsit

††††*Heteroscedasticitatea* apare atunci când varianța termenilor de eroare diferă între observații.

> qqnorm(resid(fm), main="Residuals Rankit Plot")
> *# Efectuarea unui grafic normal al scorurilor pentru a verifica asimetria, curtoza şi valorile aberante.*
> *# (Nu este foarte util aici.) Ieşire* Figura 3.9 .

Îndepărtarea acestor 5 obiecte

Verificat!

Verificat!

Verificat!

Verificat!

> manechin
Eroare: obiectul „dummy” nu a fost găsit #
Verificat!
SFÂRȘIT AL ACESTEI SESIUNI DE EXERCICE

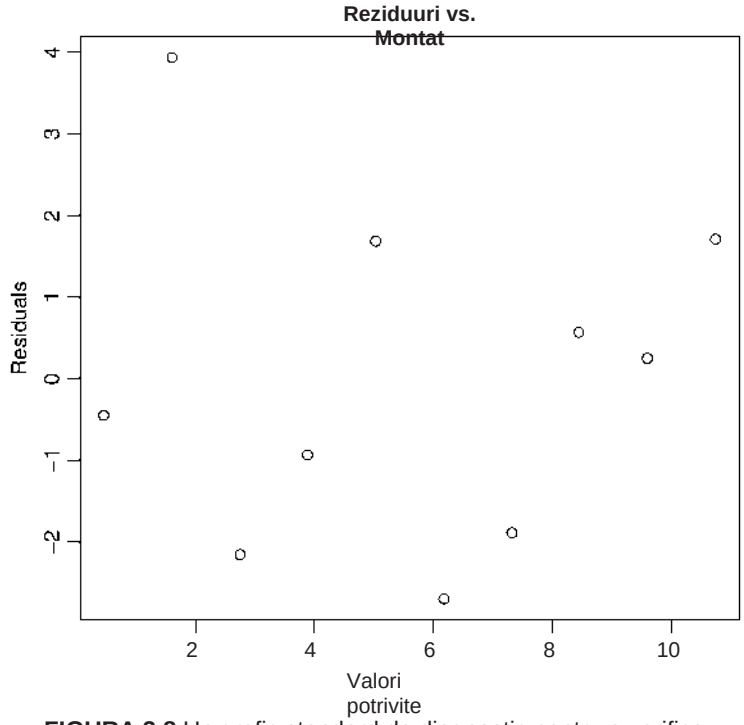


FIGURA 3.8 Un grafic standard de diagnostic pentru a verifica heteroscedasticitatea.

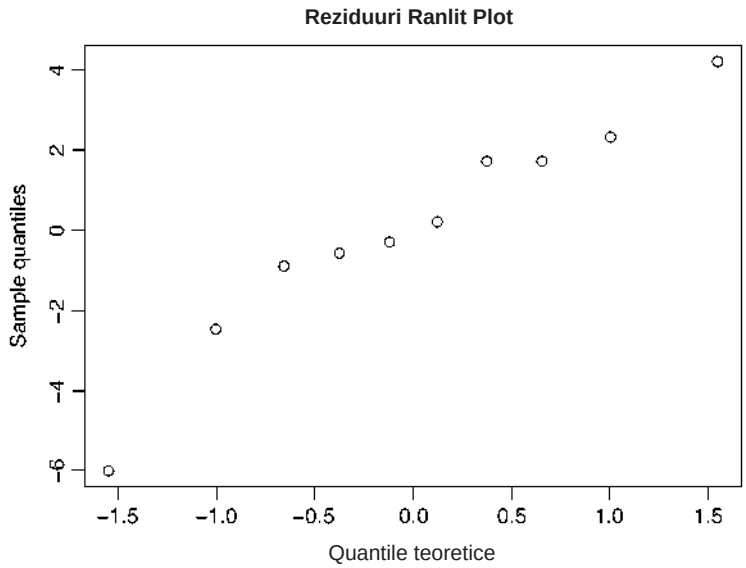


FIGURA 3.9 Un grafic normal al scorurilor pentru a verifica asimetria, curtoza și valorile aberante.

Mediul R

Trecerea prin prima sesiune (în secțiunea anterioară) arată că:

- Din punct de vedere tehnic, R este un limbaj de expresie cu o sintaxă simplă care se explică aproape de la sine. Este sensibil la majuscule și minuscule: astfel, x și X sunt simboluri diferite și se referă la variabile diferite. Sunt permise toate simbolurile alfanumerice, plus „.” și „-”, cu restricția că un nume trebuie să înceapă cu „.” sau o scrisoare. Dacă începe cu „.”, al doilea caracter *nu trebuie* să fie o cifră. Linia de comandă > indică când R este gata pentru introducere. Aici tastați comenzile care urmează să fie procesate de R ; procesarea va începe când apăsați tasta ENTER.
- Comenzile constau fie din expresii , fie din atribuții. Când o expresie este dată ca comandă, aceasta este imediat evaluată și tipărită, iar valoarea este eliminată. O atribuire evaluează o expresie și transmite valoarea unei variabile, dar valoarea *nu este* imprimată automat. Pentru a imprima valoarea calculată, pur și simplu introduceți din nou variabila la următoarea comandă.
- Comenzile sunt separate fie printr-o linie nouă, fie printr-un punct și virgulă (“;”). Mai multe comenzi elementare pot fi grupate într-o expresie compusă prin acolade (“{“ și ”}”).
- Comentariile, care încep cu un semn hash/semn numeric („#”), pot fi puse aproape oriunde. Totul până la sfârșitul rândului după acest semn este un comentariu. Comentariile nu pot fi utilizate într-o listă de argumente a unei definiții de funcție sau în interiorul șirurilor de caractere. Dacă o comandă nu este completă la sfârșitul unei linii, în mod implicit R va da un prompt diferit (un semn „+”) pe a doua și pe rândurile următoare și va continua să citească intrarea până când comanda este finalizată sintactic.
- Rezultatul unei comenzi este imprimat pe dispozitivul de ieșire . Dacă rezultatul este o matrice, cum ar fi un vector sau o matrice, atunci elementele sunt formatate cu întreruperi de linie (acolo unde este necesar) cu indicii intrărilor *principale* etichetați între paranteze drepte: [index]. De exemplu, o matrice de 15 elemente poate fi scoasă după cum urmează:

  - matrice (8, 15)

  [l] 8 8 8 8 8 8 8 8 8 8
  [ll] 8 8 8 8 8

  Etichetele [1] și [11] indică primul și al unsprezecelea element din rezultat. Aceste etichete *nu fac* parte din datele în sine.

  În mod similar, etichetele pentru o matrice sunt plasate la începutul fiecărui rând și coloană din ieșire. De exemplu, pentru matricea 3 × 5 M, aceasta este redată după cum urmează:

```
>
>  M <- matrice (1:15, nrow=3)
>  M
     [,1]  [,2]  [,3]  [,4]  [,5]
[1,]    1     4     7    10    13
[2,]    2     5     8    11    14
[3,]    3     6     9    12    15
>
```

Rețineți că stocarea este o *coloană majoră* ; adică elementele primei coloane sunt

tipărite mai întâi, urmate de cele din a doua coloană și așa mai departe. Pentru a face ca o matrice să fie completată pe rând, mai degrabă decât în mod implicit pe coloană, utilizați comutatorul suplimentar byrow=T ; acest lucru va face ca matricea să fie completată pe rând și nu pe coloană:

```
>
> M <- matrice (1:15, nrow=3, byrow=T)
>   M
     [,1]   [,2]   [,3]   [,4]   [,5]
[1,]    1      2      3      4      5
[2,]    6      7      8      9     10
[3,]   11     12     13     14     15
>
```

Prima sesiune (finalizată în subsecțiunea precedentă) arată, de asemenea, că o serie de resurse utile sunt încorporate în mediul R pe care le puteți accesa cu ușurință, folosind ajutorul online oferit de CRAN.

Întrebări de revizuire pentru Secțiunea 3.2

- Accesați fereastra „ R Graphic Device 2 (ACTIVE)".

(d) Ce este acolo?

```
> plot(z)
```

(e) Descrieți ce este în fereastra Graphic Device 2.

3.3 R CA CALCULATOR (ARAGON, 2011; DALGAARD, 2002)

Operații matematice folosind R

Pentru a învăța să faceți analize și calcule biostatistice, începeți prin a considera limbajul de programare R ca un simplu calculator. Începeți aici: introduceți doar o expresie aritmetică, apăsați tasta ENTER și căutați răspunsul de la mașină pe linia următoare.

```
>
> 2 + 3
[1]  5
>
```

Dar alte calcule? De exemplu, $13 - 7$, 3×5, $12 / 4$, 7^2, $\sqrt{2}$, e^3, $e^{i\pi}$, $\ln 5 = \log_e 5$, $(4 + \sqrt{3})(4 - \sqrt{3})$, $(4 + i\sqrt{3})(4 - i\sqrt{3})$, … și așa mai departe. Incearca doar:

```
>
> 13 - 7
[2]  6
> 3*5
[3]  15
> 12/4
[4]  3
> 7^2
[5]  49
[6] sqrt(2)
```

```
[1] 1,414214
>
[7] exp(3)
[1] 20,08554
>
> exp(1i*pi) [ 1i este folosit pentru numărul complex i = √-1.]
[1]  -1 - 0i [Aceasta este celebra ecuație de identitate a lui Euler: e^iπ + 1 = 0.]
> jurnal(5)
[2]  1,609438
> (4+sqrt(3))*(4-sqrt(3))
[3]  13 [Verificare: (4 + √3)(4 - √3) = 4^2 - (√3)^2 = 16 - 3 = 13 (Verificat!)]
> (4 + 1i*sqrt(3))*(4 - 1i*sqrt(3))
[4]  19+0i [Verificare: (4 + i√3)(4 - i√3) = 4^2 - (i√3)^2 = 16 - (-3) = 19 (Bifat!)]
```

Observație: Amintiți-vă, [1] din fața rezultatului calculat este modul lui R de a scoate numerele. Devine util atunci când rezultatul este un vector lung. Numărul *N* cuprins între paranteze [*N*] este indexul primului număr de pe acea linie. De exemplu, dacă generați 23 de numere aleatorii dintr-o distribuție normală, se obține următorul rezultat:

```
>
> x <- rnorm(23)
> x
  [1] -0,5561324  0,2478934 -0,8243522  1,0697415  1,5681899
  [6] -0,3396776 -0,7356282  0,7781117  1,2822569 -0,5413498
[11]  0,3348587 -0,6711245 -0,7789205 -1,1138432 -1,9582234
[16] -0,3193033 -0,1942829  0,4973501 -1,5363843 -0,3729301
[21]  0,5741554 -0,4651683 -0,2317168
>
```

Observație: După ce numerele aleatoare au fost generate, nu există nicio ieșire până când nu solicitați *x* ; atunci *x* devine un vector cu 23 de elemente, așa că îl numim un vector de 23.

[11] de pe a treia linie a rezultatului indică faptul că 0,3348587 (evidențiat cu gri aici pentru subliniere) este al unsprezecelea element din vectorul de 23 x . Numărul de ieșiri pe linie depinde de lungimea fiecărui element, precum și de lățimea paginii.

Atribuirea valorilor în R și calcule folosind vectori și matrici

R este conceput pentru a fi un *limbaj tip dinamic* ; adică, în orice moment, se poate schimba tipul de date al oricărei variabile. De exemplu, mai întâi puteți seta *x* să fie numeric, așa cum sa făcut până acum; să spunem, *x* = 7. Puteți, de asemenea, să setați *x* să fie un vector; să spunem, *x* = c (1, 2, 3, 4). Apoi, din nou, puteți seta *x* să fie un obiect cuvânt, cum ar fi „Bună!" Următorul mediu R :

```
>
> x <- 7
> x
[5]  7
> x <- c(1, 2, 3, 4) # x este atribuit ca fiind un 4-vector .
```

```
> x
[6]  1 2 3 4
> x <- c("Bună!") # x este atribuit ca un șir de caractere .
> x
[7]  "Hi!"
> x <- c(„Salutări și salutări!")
> x
[8]  „Salutări și salutări!"
> x <- c(„Ploaia din Spania cade în principal pe câmpie.")
[9]  „Ploaia din Spania cade în principal pe câmpie".
> x <- c(„Biostatistică", „Epidemiologie", „Sănătate publică")
> x
[10] „Biostatistică" „Epidemiologie" „Sănătate publică"
>
```

Calcule în vectori și grafice simple

Utilizarea matricelor și matricelor a fost introdusă în subsecțiunea precedentă. În matematica finită, o *matrice* este o matrice bidimensională (2D) de elemente, care sunt de obicei numere. În R , utilizarea matricei se extinde la elemente de orice tip, cum ar fi o matrice de șiruri de caractere. Matricele și matricele pot fi reprezentate ca vectori cu dimensiuni.

În biostatistică, majoritatea variabilelor poartă valori multiple, astfel încât calculele sunt de obicei efectuate între vectori ai mai multor elemente. Aceste operații între multivariate au ca rezultat matrici mari. Pentru a demonstra rezultatele, adesea sunt utile reprezentările grafice. Următorul exemplu simplu ilustrează aceste operațiuni care se realizează cu ușurință în mediul R :

```
>
[11]        greutate <- c(73, 59, 97)
[12]        înălțime <- c(1,79, 1,64, 1,73)
[13]        IMC <- greutate/înălțime^2
[14]        IMC # Citiți notele despre IMC după exemplu.
[1]  22.78331 21.93635 32.41004
>  # Pentru a rezuma rezultatele, continuați să calculați după cum urmează:
>  cbind (greutate, înălțime, IMC)
     greutate înălțime IMC
[1,]       73     1,79  22,78331
[2,]       59     1,64  21,93635
[3,]       97     1,73  32,41004
>
> rbind (greutate, înălțime, IMC)
          greutat          [,1]          [,2]          [,3]
          e          73,00000      59,00000      97,00000
          înălțim     1,79000       1,64000       1,73000
          e IMC      22.78331      21.93635      32.41004
>
```

În mod clar, funcțiile cbind și rbind leagă (unește, leagă, lipește, concatenează) vectorii pe coloană și, respectiv, pe rând, pentru a forma vectori sau matrici noi.

Utilizarea factorilor în programarea R

În analiza seturilor de date epidemiologice, sunt adesea necesare variabile categorice. Aceste variabile categoriale indică subdiviziunile setului de date inițial în diferite clase (de exemplu, vârstă, sex, stadii ale bolii, grade de diagnostic etc.). Intrarea setului de date original este, în general, delimitată în mai multe categorii folosind un cod numeric: 1 = vârstă, 2 = sex, 3 = stadiul bolii și așa mai departe. Astfel de variabile sunt specificate ca *factori* în R , rezultând o structură de date care permite cuiva să atribuie nume specifice diferitelor categorii. În anumite analize, este necesar ca R să facă distincția între codurile categoriale și variabilele ale căror valori au semnificații numerice directe.

Un factor are patru niveluri, constând din două elemente:

1. Un vector de numere întregi între 1 și 4
2. Un vector de caractere de lungime patru care conține șiruri care descriu cele patru niveluri

Luați în considerare următorul exemplu:

> Un anumit tip de cancer este clasificat în patru niveluri: stadiile 1, 2, 3 și 4.
> Nivelurile corespunzătoare de durere în concordanță cu aceste diagnostice sunt niciuna, ușoară, moderată și, respectiv, severă.
> În setul de date, cinci subiecți de caz au fost diagnosticați în ceea ce privește etapele lor respective.

Următorul segment de cod R delimitează setul de date:

```
> Cancerdurere <- c(1, 4, 3, 3, 2, 4)
> fcancerpain <- factor(cancerpain, level=1:4)
> niveluri(fcancerpain) <- c(„niciuna", „ușoară", „moderată", „severă")
```

Prima afirmație creează un vector numeric cancerpaine care codifică nivelurile de durere a șase subiecți de caz. Aceasta este considerată o variabilă categorică pentru care, folosind funcția factor , se creează un factor fcancerpain . Aceasta poate fi numită cu un singur argument în plus față de cancerpain (și anume, niveluri = 1 până la 4), ceea ce indică faptul că codarea de intrare folosește valorile 1-4. În ultima linie, numele nivelului de durere sunt schimbate în cele patru șiruri de caractere specificate. Rezultatul este:

```
> fcancerpain
[1]  niciuna severă moderată moderată ușoară severă
Nivele: niciuna ușoară moderată severă
> as.numeric(fcancerpain)
[1]  1 4 3 3 2 4
> niveluri (fcancerpain)
> 1] „niciuna" „ușoară" „moderată" „severă"
```

Observație: Funcția as.numeric scoate codificarea numerică ca numere de la 1 la 4, iar nivelurile de funcție scoate numele nivelurilor respective. Codificarea inițială de intrare în termeni de numere de la 1 la 4 nu mai este necesară. Există o opțiune suplimentară pentru a utiliza funcția ordonată , care este similară cu factorul de funcție utilizat aici.

IMC (NOTE IMC, 2012)

Indicele de masă corporală (**IMC**) este o măsură utilă pentru grăsimea corporală umană bazată pe greutatea și înălțimea unei persoane, deși *nu* măsoară de fapt procentul de grăsime din corp. Inventat la începutul secolului al XIX-lea, IMC este definit ca greutatea corporală a unei persoane (în kilograme) împărțită la pătratul înălțimii persoanei (în metri). Formula utilizată universal în medicină produce o unitate de măsură de kg/m^2:

$$\text{IMC} = \text{masa corporală (kg)}/\{\text{Înălțimea (m)}\}^2 \tag{3.3-1}$$

O diagramă IMC poate fi utilizată pentru a afișa IMC în funcție de greutate (axa orizontală) și înălțime (axa verticală), cu linii de contur pentru diferite valori ale IMC sau culori pentru diferite categorii de IMC (vezi Figura 3.10).

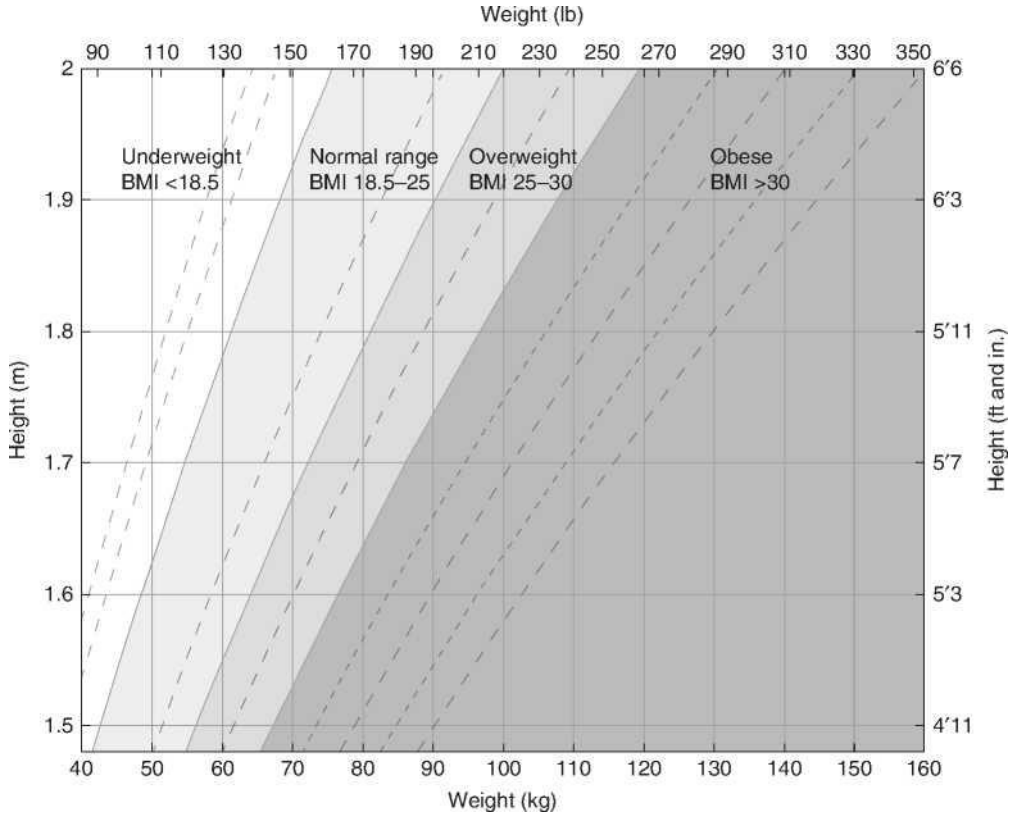


FIGURA 3.10 Un grafic al IMC. Liniile întrerupte reprezintă subdiviziuni în cadrul unei clase majore; clasificarea „Subponderală" este împărțită în continuare în subclase „severă", „moderată" și „ușoară". *Sursa:* date Organizației Mondiale a Sănătății (BMI Notes, 2012) .

Grafică simplă

Generarea de prezentări grafice este un aspect important al analizei datelor biostatistice. În mediul R , se pot construi diagrame care permit producția de grafică și controlul caracteristicilor grafice. Folosind exemplul anterior, relația dintre greutatea corporală și înălțime poate fi luată în considerare prin reprezentarea mai întâi grafică una față de alta, folosind următoarele segmente de cod R :

```
>
> parcela (greutate, inaltime)
> # Ieșire Figura 3.11 .
>
```

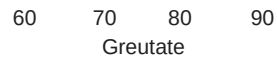


FIGURA 3.11 O diagramă xy pentru > diagramă (greutate, înălțime).

Observație: (1) Observați ordinea parametrilor din comanda plot (x, y) : primul parametru este *x* (variabila independentă, care apare pe axa orizontală) și al doilea parametru este *y* (variabila dependentă, care apare pe axa verticală).

(2) În mediul R , mulți parametri de trasare pot fi selectați pentru a modifica rezultatul. Pentru a obține o listă completă de opțiuni disponibile, reveniți la mediul R și sunați pentru:

```
> ?plot # Acesta este un apel pentru „Ajutor!” în mediul R.
> # Rezultatul este documentația R pentru:
plot {grafică} Grafic X–Y generic
```

Aceasta este documentația oficială a diagramei ***funcției*** R , în grafica ***pachetului*** R. Observați notațiile speciale folosite pentru plot și {graphics} . Pentru a utiliza pe deplin prevederile mediului R , ar trebui să investigați cu atenție toată astfel de documentație. (R are multe pachete disponibile, fiecare conținând un număr de funcții utile.) Acest document prezintă toate opțiunile de trasare disponibile cu mediul R. O copie a acestei documentații este prezentată în anexa de la sfârșitul acestui capitol pentru referință.

De exemplu, pentru a schimba simbolul de trasare, puteți utiliza cuvântul cheie pch (pentru „ caracter de tragere ”) în următoarea comandă R:

```
> plot (greutate, înălțime, pc=8)
> # Ieșirea Figura 3.12 .
```

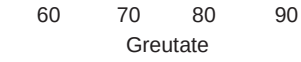


FIGURA 3.12 O diagramă xy pentru diagramă (greutate, înălțime, pch = 8).

Rețineți că rezultatul este același cu cel prezentat în Figura 3.11 , cu excepția faptului că punctele sunt marcate cu asteriscuri mici, corespunzătoare Caracterului de reprezentare pch = 8 **.**

În documentația pentru pch **,** sunt disponibile un total de 26 de opțiuni, oferind - diferite caracteristici de reprezentare a punctelor din grafica R. Ele sunt prezentate în Figura 3.13 .

```
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
                    > оЛ+хОvиЖфФ^И®ИИ»А** · О □ О Δ V
```

FIGURA 3.13 Trasarea simbolurilor în R : pch = n, n = 0, 1, 2, …, 25.

Parametrul IMC a fost ales astfel încât această valoare să fie independentă de

înălțimea unei persoane, exprimând astfel un singur număr sau indice care indică dacă un subiect de caz este supraponderal și în ce valoare relativă.

Desigur, se poate reprezenta „înălțimea" ca abscisă (axa *x orizontală*) și „greutatea" ca ordonată (axa *y verticală*), după cum urmează:

```
> plot(height, weight, pch=8) # Ieșire Figura 3.14 .
```

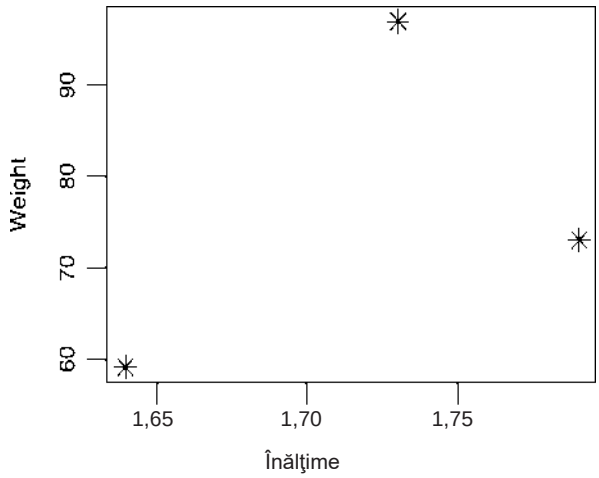


FIGURA 3.14 O diagramă xy pentru > diagramă (înălțime, greutate, pch=8).

Un IMC normal este între 18,5 și 25, cu o medie (18,5 + 25)/2 = 21,75. Pentru această valoare a IMC, atunci, greutatea unei persoane „normale" tipice ar fi (21,75 x înălțime ^{2}). Astfel, se poate suprapune o linie de greutăți „așteptate" la IMC = 21,75 pe figură 3.14. Această linie poate fi realizată în mediul R prin următoarele segmente de cod:

```
> ht <- c(1,79, 1,64, 1,73)
> lines(ht, 21.75*ht^2) # Ieşirea Figura 3.15 .
```

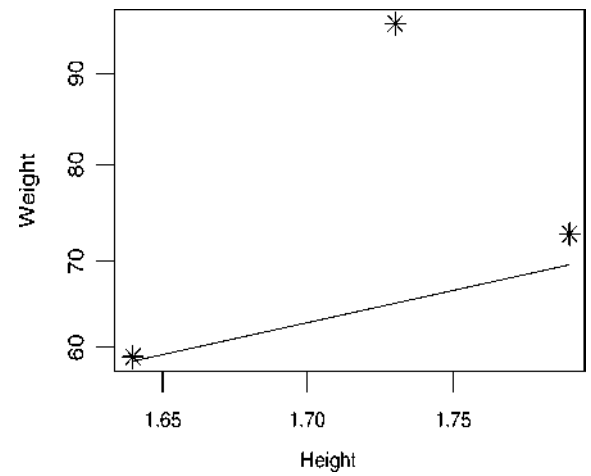


FIGURA 3.15 Curba de referință suprapusă folosind linii (ht, 21,75*ht^2).

În ultimul grafic, a fost definită o nouă variabilă pentru înălțime (ht) în loc de originală (înălțime) deoarece:

- Relația dintre înălțime și greutate este pătratică și, prin urmare, neliniară. Deși s-ar putea să nu fie evident pe diagramă, este de preferat să folosiți puncte care sunt distribuite uniform de-a lungul axei *x* , decât să vă bazați pe distribuția datelor originale.
- Deoarece valorile înălțimii nu sunt sortate, segmentele de linie nu ar conecta - punctele învecinate, ci ar rula înainte și înapoi între punctele îndepărtate.

Observatii:

1. În ultimul dintre exemplele precedente **,** R făcea de fapt aritmetica vectorilor.
2. Observați că cei doi vectori greutatea și înălțimea sunt ambii 3 vectori, ceea ce face rezonabil să efectuați următorul pas.
3. Instrucțiunea cbind , când este utilizată imediat după ce calculele au fost finalizate, formează o nouă matrice *legând* împreună matrice pe orizontală sau *pe* coloane. Rezultă o variabilă de răspuns multivariată. În mod similar, instrucțiunea rbind face o operație similară pe verticală sau *pe* rând.
4. Dacă dintr-un motiv oarecare (cum ar fi o greșeală într-una dintre intrări) cele două intrări greutate și înălțime au un număr *diferit* de elemente, R va afișa un mesaj de eroare. De exemplu:

```
>
> greutate <- c(73, 59, 97)              # un 3-vector
>  înălțime <- c(1,79, 1,64, 1,73, 1,48) # un vector cu 4
>  bmi <- greutate/înălțime^2 # Ieşire:
```

Mesaj de avertizare: *# Un mesaj de eroare!*
În greutate/înălțime^2:
lungimea mai mare a obiectului nu este un multiplu al lungimii mai scurte a obiectului

```
>
```

x ca Vectori și Matrici în Biostatistică

Tocmai am văzut că o variabilă, cum ar fi x sau M , poate fi atribuită după cum urmează:

1. Un număr, cum ar fi $x = 7$
2. Un vector sau o matrice, cum ar fi *x* = c(1, 2, 3, 4)
3. O matrice, cum ar fi *x* =

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1	4	7	10	13
[2,]	2	5	8	11	14
[3,]	3	6	9	12	15

4. Un șir de caractere, cum ar fi
   x = „Ploaia din Spania cade în principal pe câmpie".
5. De fapt, în R , unei variabile *x* i se poate atribui un set de date complet, care poate consta dintr-un set multidimensional de elemente, fiecare dintre acestea putând fi, la rândul său, oricare dintre aceste tipuri de variabile. De exemplu, pe lângă faptul că este un vector numeric, ca în numărul 2 din această listă, *x* poate fi:
   (a) un *vector de caractere* , care este un vector de șiruri de text ale căror elemente sunt exprimate prin ghilimele, folosind ghilimele duble, simple sau mixte:

```
> c(„unu", „doi", „trei", „patru",  „cinci") # Ghilimele
                                    "cinci"
[1] „unu" „doi" „trei" „patru" >
                                    „cinci") # ghilimele
> c(„unu", „doi", „trei", „patru",  simple „cinci"
[1] „unu" „doi" „trei" „patru" >
> c(„unu", „doi", „trei", „patru",  „cinci") # Citate mixte
[1] „unu" „doi" „trei" „patru"      „cinci"
```

   Cu toate acestea, o pereche mixtă de ghilimele, cum ar fi „xxxxx" , *nu va* fi acceptată. De exemplu:

```
> c(„unu", „doi", „trei", „patru", „cinci")
+
```

   (b) un *vector logic*, care ia valoarea TRUE sau FALSE (sau NA). Pentru intrări, se poate folosi abrevierea T sau F . Acești vectori sunt specificați în mod similar folosind funcția c :

```
> c(T, F, T, F, T)
[1] ADEVĂRAT FALS ADEVĂRAT FALS ADEVĂRAT
```

   În cele mai multe cazuri, nu este nevoie să repetați vectorii logici specificați. Este acceptabil să se folosească o singură valoare logică pentru a oferi opțiunile necesare, deoarece vectorii cu mai multe valori vor răspunde în termeni de expresii relaționale . Observati:

```
> greutate <- c(73, 59, 97)
> înălțime <- c(1,79, 1,64, 1,73)
> IMC <- greutate/înălțime^2
```

```
>  bmi # Ieșire:
[1] 22,78331 21,93635 32,41004
>  bmi > 25 # O singură valoare logică va fi suficientă.
[1]  FALS FALS ADEVĂRAT
>
```

Câteva funcții speciale care creează vectori

Trei funcții care creează vectori sunt c, seq și rep.

c , pentru „concatena"; îmbinarea obiectelor cap la cap (aceasta a fost introdusă mai devreme). De exemplu:

```
> x <- c(1, 2, 3, 4) # x este atribuit ca fiind un 4-vector.
> x
[1]  1 2 3 4
```

seq , pentru „secvență"; definind o succesiune echidistantă de numere. De exemplu:

```
> seq(1, 20, 2) # Pentru a scoate o secvență de la 1 la 20 în pași de 2.
[1]  1 3 5 7 9 11 13 15 17 19
>  seq(1, 20)   # Pentru a scoate o secvență de la 1 la 20 în pași de 1 (care poate
>               # fie omis).
[1]  1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
> 1:20 # Aceasta este o alternativă simplificată la scrierea seq(1, 20) .
[1]  1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
> seq(1, 20, 2.5) # Pentru a scoate o secvență de la 1 la 20 în pași de 2,5.
[1]  1,0 3,5 6,0 8,5 11,0 13,5 16,0 18,5
```

rep , pentru „replica" ; pentru generarea de valori repetate. Această funcție are două forme, în funcție de faptul că al doilea argument este un singur număr sau un vector. De exemplu:

```
>  rep(1:2, c(3,5)) # Replicarea primului element (1) de 3 ori și
>                   # apoi replicând al doilea element (2) de 5 ori
>  1] 1 1 1 2 2 2 2 2 # Aceasta este ieșirea.
>  vector <- c(1, 2, 3, 4)
>  vector # Vector de ieșire
[1]  1 2 3 4
> rep(vector, 5) # Replicarea vectorului de 5 ori
[1] 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4
```

Matrice și matrice

În matematica finită, o matrice M este o matrice 2D de elemente (în general numere), cum ar fi:

$$M = \begin{matrix} 1 & 4 & 7 & 10 & 13 \\ 2 & 5 & 8 & 11 & 14 \\ 3 & 6 & 9 & 12 & 15 \end{matrix}$$

Matricea este de obicei plasată în paranteze () sau în unele tipuri de paranteze {}, []. În R , utilizarea unei matrice este extinsă la elemente de mai multe tipuri, numere, precum și șiruri de caractere. De exemplu, în R , matricea eșantionului precedentă M este

exprimată după cum urmează:

```
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    4    7   10   13
[2,]    2    5    8   11   14
[3,]    3    6    9   12   15
```

Utilizarea funcției de dimensiune dim în R

În R , matricea 3 × 5 eșantion precedentă poate fi configurată ca vectori cu dimensiunea dim(x) folosind următorul segment de cod:

```
> x <- 1:15
> x
>  1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
>   dim(x) <- c(3, 5) # O dimensiune de 3 rânduri pe 5 coloane
>   x
      [,1] [,2] [,3] [,4] [,5]
[1,]    1    4    7   10   13
[2,]    2    5    8   11   14
[3,]    3    6    9   12   15
```

Observație: Aici 15 elemente totale, de la 1 la 15, sunt setate să fie elementele matricei *x* . Apoi dimensiunea lui *x* este stabilită ca c(3, 5), făcând din *x* o matrice 3 × 5 . Alocarea celor 15 elemente urmează o procedură pe coloană, astfel încât elementele primei coloane sunt alocate mai întâi, urmate de cele ale coloanei a doua, apoi celei de-a treia coloane și așa mai departe.

Utilizarea matricei Funcții Matrice în R

O altă modalitate de a genera o matrice este să utilizați funcția matrice . Matricea 3 × 5 utilizată în subsecțiunea anterioară poate fi creată de următorul segment de cod cu o singură linie:

```
>   matrice (1:15, nrow=3)
      [,1] [,2] [,3] [,4] [,5]
[1,]    1    4    7   10   13
[2,]    2    5    8   11   14
[3,]    3    6    9   12   15
```

Cu toate acestea, dacă cele 15 elemente ar trebui alocate pe rând, atunci trebuie utilizat următorul segment de cod:

```
>   matrice (1:15, nrow=3, byrow=T)
      [,1] [,2] [,3] [,4] [,5]
[1,]    1    2    3    4    5
[2,]    6    7    8    9   10
[3,]   11   12   13   14   15
```

Câteva funcții utile care operează pe matrici în R

> colnames , rownames și t (pentru *transpunere*)

Folosind același exemplu de matrice 3 × 5 , mai întâi celor cinci coloane ale matricei 3 × 5 *x* li se atribuie numele C1, C2, C3, C4 și, respectiv, C5. Apoi se obține transpunerea și în final se ia transpunerea transpunerii pentru a obține matricea originală *x* :

```
>  matrice (1:15, nrow=3, byrow=T)
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    2    3    4    5
[2,]    6    7    8    9   10
[3,]   11   12   13   14   15
>  colnames(x) <- c("C1", "C2", "C3", "C4", "C5")
> x
     C1   C2   C3   C4   C5
[1,]    1    4    7   10   13
[2,]    2    5    8   11   14
[3,]    3    6    9   12   15
>  t(x)
     [,1] [,2] [,3]
C1      1    2    3
C2      4    5    6
C3      7    8    9
C4     10   11   12
C5     13   14   15
>  t(t(x)) # care este doar x, așa cum era de așteptat.
     C1   C2   C3   C4   C5
[1,]    1    4    7   10   13
[2,]    2    5    8   11   14
[3,]    3    6    9   12   15
```

O altă modalitate de a face acest lucru este să utilizați funcția LETTERS , care este o variabilă încorporată care conține literele majuscule de la A la Z. Alți vectori utili includ litere, month.name și month.abb pentru litere mici, nume de luni, și, respectiv, nume prescurtate de luni. Aruncă o privire:

```
>  X <-LITERE
>  X
[1]  "A" "B" "C" "D" "E" "F" "G" "H" "I" "J" "K" "L" "M" "N" "O"
[16]  "P" "Q" "R" "S" "T" "U" "V" "W" "X" "Y" "Z"
> x <-litere
> x
[1]  "a" "b" "c" "d" "e" "f" "g" "h" "i" "j" "k" "l" "m" "n" "o"
[16]  "p" "q" "r" "s" "t" "u" "v" "w" "x" "y" "z"
> M <- luna.nume
> M
[I]   „ianuarie” „februarie” „martie” „aprilie”                    „mai”
[6] „lunie” „iulie”                  „august” „septembrie” „octombrie”
[II]   „noiembrie” „decembrie”
[III]  m <- luna.abb
> m
[I]   „Ian” „Feb” „Mar” „Apr” „Mai” „Iun” „Iul” „Aug” „Sep” „Oct”
[II]   "noiembrie" "dec."
```

NA: „Nu este disponibil” pentru valorile lipsă din seturile de date

NA este o constantă logică de lungime 1, care conține un indicator de valoare lipsă.

NA poate fi forțat la orice alt tip de vector, cu excepția brutului. Există, de asemenea, constante NA_integer_, NA_real_, NA_complex_ și NA_character_ ale celorlalte tipuri de vectori atomici care acceptă valorile lipsă. Toate acestea sunt cuvinte rezervate în limbajul R.

- Funcția generică .na indică elementele care lipsesc.
- Funcția generică .na<- setează elementele la NA .

Cuvintele rezervate din analizatorul lui R sunt if , else , repeat , while , function , for , next , break , NA_complex _, NA_character _, ..., and ... 1 , ... 2 , și așa mai departe, care sunt folosit pentru a se referi la argumente transmise dintr-o funcție de încadrare.

Cuvintele rezervate din afara ghilimelelor sunt întotdeauna analizate pentru a fi referințe la obiectele la care se face legătura în lista de mai sus și nu sunt permise ca nume sintactice. Sunt *permise* ca nume nonsintactice.

Funcții speciale care creează vectori

Există trei funcții R utile care sunt adesea folosite pentru a crea vectori:

1. c pentru „concatenare”, care a fost introdus mai devreme în această secțiune pentru unirea elementelor cap la cap. De exemplu:

   ```
   > c(2, 3, 5, 7, 11, 13, 17, 19, 23, 29) # Primele 10 numere prime
   [1] 2 3 5 7 11 13 17 19 23 29
   ```

2. seq pentru „secvență”, care este folosit pentru a enumera secvențe echidistante de numere. De exemplu:

   ```
   > seq(1, 20) # Secvență de la 1 la 20
   [1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
   > seq(1, 20, 1) # Secvență de la 1 la 20 în pași de 1
   [1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
   > 1:20 # Secvență de la 1 la 20 în pași de 1
   [1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
   > seq(1, 20, 2) # Secvență de la 1 la 20 în pași de 2
   [1] 1 3 5 7 9 11 13 15 17 19
   > seq(1, 20, 3) # Secvență de la 1 la 20 în pași de 3
   [1] 1 4 7 10 13 16 19

   > seq(1,20, 10) # Secvență de la 1 la 20 în pași de 10
   [1] 1 11
   > seq(1,20, 20) # Secvență de la 1 la 20 în pași de 20
   [1] 1
   > seq(1,20, 21) # Secvență de la 1 la 20 în pași de 21
   [1] 1 >
   ```

3. rep pentru „replicat”, care este folosit pentru a genera valori repetate și poate fi exprimat în două moduri. De exemplu:

```
>  x <- c(3, 4, 5)
>  rep(x, 4)          # Replica vectorul x de 4 ori.
[1] 3 4 5 3 4 5 3 4 5 3 4 5
>  rep(x, 1:3)        # Replica elementele lui x: primul element o dată, al doilea
>                     # element de două ori, iar al treilea element de trei ori.
[1] 3 4 4 5 5 5
>  rep(1:3, c(3,4,5)) # Pentru secvența (1,2,3), replicați elementele sale 3,
>  # 4 și, respectiv, de 5 ori.
[1]  1 1 1 2 2 2 2 3 3 3 3 3
```

Întrebări de revizuire pentru Secțiunea 3.3

1. Generați un „Turn al puterilor” prin calcule folosind R . Există o provocare interesantă în aritmetică, care arată astfel:

$$\sqrt{2}^{\sqrt{2}^{\cdots}}$$

Care este valoarea lui $\sqrt{2}^{\sqrt{2}^{\cdots}}$? Aceasta este o infinitate de „turn al puterilor” ascendent al rădăcinii pătrate a lui 2.

Soluție: Fie *x* valoarea acestui „Turn al Puterilor”. Atunci se vede cu ușurință că $\sqrt{2}^{x} = x$ însuși. Sunteți de acord? Urmăriți cel mai mic $\sqrt{2}$. În mod clar, rezultă că $x = 2$, deoarece $\sqrt{2}^{2} = 2$. Aceasta arată că valoarea acestui Turn infinit al puterilor de $\sqrt{2}$ este doar 2.

Acum utilizați mediul R pentru a verifica acest rezultat interesant:

(a) Calculați $\sqrt{2}$

```
> sqrt(2)
```

(b) Calculați $\sqrt{2}^{\sqrt{2}}$

```
> sqrt(2)^sqrt(2) [a 2-Turnuri de √2-s]
```

(c)
```
> sqrt(2)^sqrt(2)^sqrt(2) [a 3-Turnuri de √2-s]
```

(d)
```
> sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2) [a 4-Turnuri de √2-s]
```

(e)
```
> sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2) [a 5-Turnuri de √2-s]
```

(f) Acum, încercați următoarele calcule de 10-, 20-, 30- și, în final, 40-Towers of Powers of $\sqrt{2}$ și, în final, ajungeți la rezultatul 2 (cu exactitate cu șase zecimale). >

```
sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^
sqrt(2)
[1]  1,983668 [un turn cu 10 puteri de √2 -s]
> sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^s
qrt(2) )^
+ sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^s
qrt(2) )
[1]  1,999586 [un turn de 20 de puteri de √2-s]
> sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^s
qrt(2) )^
+ sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^s
qrt(2) )^
+ sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^s
```

```
qrt(2) )
[2]  1,999989 [un turn de 30 de puteri de √ 2-s]
> sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^s
qrt(2) )^
+ sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^s
qrt(2) )^
+ sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^s
qrt(2) )^
+ sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^s
qrt(2) )
[3]  2 [un turn de 40 de puteri de √ 2-s]
```

Astfel, acest calcul R verifică soluția.

2. (a) Care sunt echivalentele în R pentru operațiile matematice de bază : +, -, × , / (diviziunea), √ și pătratul unui număr?
   (b) Descrieți utilizarea factorilor în programarea R. Da un exemplu.
3. Dacă x = (0, 1, 2, 3, 4, 5) și y = (0, 1, 4, 9, 16, 25), utilizați R pentru a reprezenta grafic:
   (a) y față de x
   (b) x față de y
   (c) $\sqrt{y}$ față de x
   (d) y față de $\sqrt{x}$
   (e) $\sqrt{y}$ față de $\sqrt{x}$
   (f) $\sqrt{x}$ față de $\sqrt{y}$
4. Explicați, dând un exemplu, cum pot fi utilizate următoarele funcții pentru a combina matrice pentru a forma altele noi: (a) cbind , (b) rbind .
5. (a) Descrieți factorul funcției R () .
   (b) Dați un exemplu de utilizare a factorului() pentru a crea matrice noi.
6. Folosind exemple, ilustrați două proceduri de creare:
   (a) un vector
   (b) o matrice
7. Descrieți, folosind exemple, următoarele trei funcții pentru crearea vectorilor:
   (a) c
   (b) secv
   (c) reprezentant
8. (a) Utilizați funcția dim() pentru a configura o matrice. Da un exemplu.
   (b) Utilizați funcția matrix() pentru a configura o matrice. Da un exemplu.

9. Descrieți, folosind un exemplu, utilizarea următoarelor funcții care operează pe o matrix în R: t(), colnames() și rownames() .
10. (a) Care sunt cuvintele rezervate în mediul R ?
(b) În R , cum este utilizată constanta logică NA ? Da un exemplu.

Exerciții pentru Secțiunea 3.3 (Everitt & Hothorn, 2006; Virasakdi, 2011)

Intrați în mediul R și faceți următoarele exerciții folosind programarea R.

(a) Construiți o matrice care să arate IMC-urile lor ca ultimul rând.
(b) Intreg:
(i) GREUTATE (pe axa *y*) față de ÎNĂLȚIE (pe axa *x*)
(ii) ÎNĂLȚIE versus GREUTATE
(iii) Presupunând că greutatea unei persoane „normale" tipice este (21,75 × ÎNĂLȚIE [15]), suprapuneți o linie de greutate „așteptată" la IMC = 21,75 pe graficul de la (i).

4. (a) Pentru a converti temperaturile în grade Fahrenheit (F) și Celsius (C), se folosesc următoarele formule de conversie:

$$F = (9/5)C + 32$$

$$C = (5/9) \times (F - 32)$$

3. La temperatura și presiunea standard, punctele de îngheț și de fierbere ale apei sunt de 0, respectiv 100 de grade Celsius. Care sunt punctele de îngheț și de fierbere ale apei în grade Fahrenheit?

	John	Chang	Michael	Bryan	Jose
WEIGHT	69.1	62.5	74.3	70.9	96.6
HEIGHT	1.81	1.46	1.69	1.82	1.74

(b) Pentru C = 0, 5, 10, 15, 20, 25, ..., 80, 85, 90, 95, 100, calculați un tabel de conversie care arată temperaturile Fahrenheit corespunzătoare.

Notă: Pentru a crea secvența temperaturilor Celsius, utilizați funcția R seq(0, 100, 5) .

5. Utilizați datele din Tabelul 3.1 (Aragon, 2011 ; CDC, 2005) . Să presupunem că o persoană este inițial HIV-negativă.

Dacă probabilitatea de a se infecta pe act este *p* , atunci probabilitatea de a *nu* se infecta pe act este (1 - *p*).

Probabilitatea de *a nu* se infecta după două acte consecutive este $(1 - p)^2$, iar probabilitatea de a *nu* se infecta după trei acte consecutive este $(1 - p)^3$.

Prin urmare, probabilitatea de a nu se infecta după *n* acte consecutive este $(1 - p)^n$, iar probabilitatea de a se *infecta* după *n* acte consecutive este $1 - (1 - p)^n$.

(a) Pentru probabilitatea de transmitere fără transfuzie de sânge (risc pe act) din

15IMC este calculat din greutatea în kilograme și înălțimea în metri:

$$IMC = kg/m^2$$

Folosind 1 kg ≈ 2,2 lb și 1 m ≈ 3,3 ft ≈ 39,4 in.:
(a) Calculați-vă IMC.
(b) Este în intervalul „normal" 18,5 ≤ IMC ≤ 25?

tabel 3.1, se calculează riscul de a fi infectat după 1 an (365 de zile) dacă se efectuează consumul de droguri injectabile (IDU) o dată pe zi timp de 1 an.

(b) Aceste riscuri cumulate par rezonabile? De ce sau de ce nu?

TABELUL 3.1 Riscul estimat per-act (probabilitatea de transmitere) pentru dobândirea HIV pe calea de expunere la o sursă infectată

RISCUL	CALEI DE EXPUNERE LA 10.000 DE EXPUNERI
Transfuzie de sânge (BT)	9.000
Consumul de droguri injectabile (IDU)	67

Sursa: CDC, 2005

SOLUŢIE:

```
> p <-67/10000
> P
[1] 0,0067
> q <- (1 - p)
> q
[1] 0,9933
> q365 <- q^365
> q365
[1] 0,08597238
> p365 <- 1 - q365
> p365
[1] 0,9140276
```

= > Probabilitatea de a fi infectat într-un an = 91,40%. Un risc mare, într-adevăr!

3.4 UTILIZAREA R ÎN ANALIZA DATELOR ÎN BIOS

concluzii și probabil recomandări pentru cursuri definitive de acțiuni pentru îmbunătățirea sănătății publice și personale. În această secțiune sunt examinate mai multe metode de introducere a setului de date .

Introducerea datelor la linia de comandă R

CADRE DE DATE ȘI SETURI DE DATE (VIRASAKDI, 2011)

Mulți investigatori epidemiologici folosesc termenii *cadru de date* și *set de date în mod* interschimbabil . Cu toate acestea, se pot face distincții.

În multe aplicații, un set complet **de date** conține mai multe cadre de date, inclusiv datele reale care au fost colectate.

Regulile pentru **cadrele de date** sunt similare cu cele pentru matrice și matrice, introduse mai devreme în Secțiunea 3.3. Cu toate acestea, cadrele de date sunt mai complicate decât matricele. Într-o matrice, dacă o singură celulă este un caracter, atunci toate coloanele vor fi caractere. În schimb , un cadru de date poate consta din:

- O coloană „IDnumber”, în care datele sunt numerice
- O coloană „Nume”, în care datele sunt caractere

Într-un cadru de date, fiecare variabilă poate avea descrieri lungi ale variabilelor, iar un factor poate avea „niveluri" sau niveluri de valoare. Aceste proprietăți pot fi transferate din setul de date original în alte formate de software (cum ar fi SPS, Stata etc.). Ele pot fi create și în R .

CREAREA UNUI CADRU DE DATE PENTRU COMPUTAREA *R* FOLOSIND O Foaie DE TACUL EXCEL (PlatFORMA WIN DOWS)

Ca exemplu folosind un set tipic de date reale de cercetare epidemiologică **caz-control** , luați în considerare setul de date din Tabelul 3.2 . Aceste date au fost extrase dintr-un studiu clinic pentru a evalua eficacitatea chimioterapiei de întreținere pentru subiecții de caz cu leucemie mielogenă acută (AML), efectuat la Universitatea Stanford, California, în 1977. După ce a ajuns la un statut de remisie prin tratament prin chimioterapie, pacienții care au intrat în studiu au fost repartizați *aleatoriu* într-unul din cele două grupuri:

(1) Menținut: acest grup a primit chimioterapie de întreținere.
(0) Nemenținut: acest grup nu a primit chimioterapie; este grupul de control.

Studiul clinic a fost efectuat pentru a stabili dacă chimioterapia de întreținere a prelungit timpul până la recidivă (=„moarte").

Vom folosi următoarea procedură (a) pentru *a crea* un fișier de date AML, numit AML. csv, în Windows; și (b) *pentru a introduce* noul fișier de date în R ca fișier de date AML .

Crearea unui cadru de date pentru calculul R:

1. Introducerea datelor, folosind Microsoft Excel:
   (a) Deschide foaia de calcul Excel.
   (b) Introduceți datele astfel încât numele variabilelor să fie în rândul 1 al foii de calcul Excel. (c) Luați în considerare fiecare rând de date pentru a reprezenta un subiect de caz individual din studiu. (d) Începeți cu coloana A.

TABEL 3.2 Date pentru studiul clinic de întreținere AML*

GRUP	A + INDICĂ O VALOARE CENZURĂ DURATA REMISIEI COMPLETE (SĂPTĂMÂNI)	
1 = menținut (11)	9,13,13+,18,23,28+,31,34,45+,48,161+ }	1 = Necenzurat
0 = Neîntreținut (12)	5, 5, 8, 8, 12, 16+,23,27,30,33,43,45}	0 = cenzurat (+)

NB: Grupul nemenținut poate fi considerat MBD.**

*Puncte de date preluate din *Survival Analysis Using S: Analysis of Time-to-Event Data* , de Mara Tableman și Jong Sung Kim (Boca Raton, FL: Chapman & Hall/CRC, 2004).

- *Epigenomul cancerului este caracterizat prin modele specifice de metilare ADN și modificare a cromatinei. Proteinele care mediază aceste modificări sunt codificate de genele epigenetice definite aici ca:
- ADN metiltransferaze (DNMT)
- din domeniul de legare a metil-CpG (MBD).
- histon acetiltransferaze (HAT)
- histon deacetilaze (HDAC)
- histon metiltransferaze (HMT)
- histon demetilazele

2. Salvați foaia de calcul ca fișier .csv:

(a) *Faceți clic pe* : „Fișier” → „Salvare ca” → și apoi, în caseta cu numele fișierului (caseta de sus din jos) tastați: AML.
(b) În caseta „Salvează în:” (în partea de sus), *alege* : „Disc local (C:)”. Fișierul AML va fi apoi salvat în nivelul superior al unității C; puteți alege un alt nivel sau locație dacă doriți.
(c) În caseta „Salvare ca tip” (caseta de jos din partea de jos), *derulați în jos, selectați și faceți clic* pe: CSV (Delimitat prin virgulă = Valori separate prin virgulă).
(d) Pentru a închide Excel, *faceți clic pe* „X” mare din colțul din dreapta sus.

3. În Windows, verificați unitatea C: pentru fișierul AML.csv.
4. Citiți AML în R :
   (a) Deschide R.
   (b) Utilizați funcția read.csv() :
   > aml <- read.csv("C:\\AML.csv", antet = T,sep= ",").
   (c) Acest lucru se poate face și prin:
   > aml <- read.csv(„C:\\AML.csv")
   > *# Citiți fișierul* AML.csv *de pe unitatea C: a computerului și apelați-l*
   > *# aml* **.**
5. Ieșiți fișierul AML.csv pentru inspecție:
   > aml *# Ieșire:*

	săptămân	grup	stare
1	9	1	1
2	13	1	1
3	13	1	0
4	18	1	1
5	23	1	1
6	28	1	0
7	31	1	1
8	34	1	1
9	45	1	0
10	48	1	1
11	161	1	0
12	5	0	1
13	5	0	1
14	8	0	1
15	8	0	1
16	12	0	1
17	16	0	0
18	23	0	1
19	27	0	1
20	30	0	1
21	33	0	1
22	43	0	1
23	45	0	1

>

Mai târziu în această carte, în Secțiunea 7.3 din Capitolul 7 , acest set de date va fi revizuit și procesat în continuare pentru analiza supraviețuirii.

OBȚINEREA UNUI CADRU DE DATE DIN UN FIȘIER TEXT

Datele din diverse surse sunt adesea introduse folosind multe programe software *diferite* . Acestea pot fi transferate dintr-un format în altul prin formatul de fișier ASCII. De exemplu, în Windows, un fișier text este cel mai comun fișier ASCII, având de obicei extensia „.txt”. Există și alte fișiere în format ASCII, inclusiv fișierul de comandă „ **.R** ”.

Datele din majoritatea programelor software pot fi ieșite sau salvate ca fișier ASCII. Din Excel, un program de calcul foarte popular, datele pot fi salvate în format „.csv” (valori separate prin virgulă). Aceasta este o modalitate ușoară de interfață între fișierele de foi de calcul Excel și R . Deschideți fișierul Excel și „salvați ca” formatul .csv.

Fișiere cu separatori de câmpuri: de exemplu, să presupunem că fișierul **csv1.xls** a fost inițial o foaie de calcul Excel. După ce este salvat în format .csv, fișierul de ieșire se numește **csv1.csv** , al cărui conținut este:

```
"nume","sex","varsta"
   „A”,      „F”,   20
   „B”,      „M”,   30
   „C”,      „F”,   40
```

Caracterele sunt cuprinse între ghilimele, iar *delimitatorii* (separatorii de variabile) sunt virgule. Uneori, un fișier poate să nu conțină ghilimele, ca în fișierul **csv2.csv** :

```
nume,    gen,        vârstă
   O,          F,          20
   B,        M,          30
   C,          F,          40
```

Pentru ambele fișiere, comanda R de citit în setul de date este aceeași:

```
> a <- read.csv(„csv1.csv”, as.is=TRUE)
> a
```

	numele	**genul F**	**vârst ă**
1	**A**		
2	**B**	**M**	**30**
3	**C**	**F**	**40**

Argumentul as.is=TRUE păstrează toate caracterele așa cum sunt; altfel, personajele ar fi fost forțate în factori. Variabila „nume” nu trebuie luată în considerare, dar „gen” ar trebui. Prin urmare, următoarea comandă trebuie introdusă după cum urmează:

```
>   a$gen <- factor(a$gen)
```

Rețineți că obiectul a are un cadru de date de clasă și că numele variabilelor din cadrul de date „ a ” trebuie să fie referite folosind notația semnul dolar $. În caz contrar, R va afirma că obiectul „gen” nu poate fi găsit.

Pentru fișierele cu spațiu alb (spații și file) ca separator, cum ar fi fișierul **data1.txt** , comanda de utilizat este read.table() :

```
>   a <- read.table(„data1.txt”, header=TRUE, as.is=TRUE)
```

Fișiere fără separatori de câmp: luați în considerare fișierul **data2.txt** , care este într-un format de câmp fix, *fără* separatori de câmp.

```
    nume sex varsta
1      A      F      20
2      B      M      30
```

3 CF 40

Pentru a citi într-un astfel de fișier, utilizați funcția read.fwf() :

1. Sari peste prima linie, care este antetul.
2. Trebuie specificate lățimea fiecărei variabile și numele coloanelor:

```
> a <- read.fwf("data2.txt", skip=1, width=c(1,1,2), col.names
+                    = c(„nume", „sex", „vârstă"), as.is=TRUE)
```

INTRODUCEREA ȘI ANALIZA DATELOR FOLOSIND FUNCȚIA DATA.ENTRY()

Secțiunea anterioară s-a ocupat de crearea cadrelor de date prin citirea datelor create din programe din afara lui R , cum ar fi Excel. De asemenea, este posibil să introduceți date direct în R folosind funcția data.entry() . Cu toate acestea, dacă cantitatea de date este mare (să zicem, mai mult de 15 coloane și/sau mai mult de 25 de rânduri), șansa de eroare umană este mare la introducerea datelor în foaie de calcul sau în modul text. Un program software special conceput pentru introducerea datelor, cum ar fi *EpiData* (www.epidata.dk) , este mai potrivit.

INTRODUCEREA DATELOR FOLOSIND MAI MULTE FUNCȚII R DISPONIBILE

Setul de date din tabelul 3.3 (Aragon, 2011) , care enumerează decesele în rândul subiecților cărora li s-a administrat o doză de tolbutamidă sau un placebo în Programul de Diabet al Grupului Universității (1970), este stratificat în funcție de vârstă.

TABEL 3.3 Decese în rândul subiecților cărora li s-a administrat tolbutamidă sau un placebo în cadrul programului de diabet al grupului universitar (1970)*

	VARSTA < 55		VARSTA ≥ 55		COMBINAT	
	TOLBUTAMIDE	PLACEBO	TOLBUTAMIDE	PLACEBO	TOLBUTAMIDE	PLACEBO
Decese	8	5	22	16	30	21
Supraviețui	98	115	76	69	174	184

*Disponibil la http://www.medepi.net/data/ugdp.txt

Funcțiile R care pot fi utilizate pentru a importa cadrul de date au fost introduse în Secțiunea 3.3, mai devreme în acest capitol.

O modalitate convenabilă de a introduce date la promptul de comandă este să utilizați funcțiile R c(), matrix(), array(), apply(), list(), data.frame() și odd.ratio() , ca prezentate de următoarele exemple, care utilizează datele din tabelul 3.3 :

```
>  #Introducerea datelor pentru un vector
>  vector1 <- c(8, 98, 5, 115) # Folosind datele din Tabelul 3.3 .
>  vector1
[1]   8 98 5 115
>
> vector2 <-c(22, 76, 16, 69); vector2 # Date din Tabelul 3.3 .
>  1] 22 76 16 69
>
```

```
>  # Introducerea datelor pentru o matrice
>  matrice1 <- matrice(vector1, 2, 2)
>  matricea 1

     [,1] [,2]
[1,]    8    5
[2,] 98 115

>  matrice2 <- matrice(vector2, 2, 2); matricea 2
     [,1] [,2]
[1,] 22      16
[2,] 76      69
>

>  # Introducerea datelor pentru o matrice
>  udata <- array(c(vector1, vector2), c(2, 2, 2))
> udata

, , 1
     [,1] [,2]
[1,]    8    5
[2,] 98 115

, , 2
     [,1] [,2]
[1,] 22      16
[2,] 76      69

>  aplica(udata, c(1, 2), suma); udata.tot
     [,1] [,2]
[1,] 30      21
[2,] 174 184
>

>  # Introducerea unei liste
>  x <- list(crude.data = udata.tot, stratified.data = udata)
>  x$crude.date

     [,1] [,2]
[1,] 30      21
[2,] 174 184
>  x$stratificat

>  , 1
     [,1] [,2]
[1,]    8    5
[2,] 98 115

>  , 2
     [,1] [,2]
```

[1,] 22 16
[2,] 76 69
>

> *# Introducerea unui cadru de date simplu*
> nume subiect <- c ("Petru", "Pavel", "Maria")
> subiectnumber <- 1:lungime(nume subiect)
> vârsta <- c(26, 27, 28) *# Acestea sunt adevăratele vârste ale cântăreților, respectiv, în 1964* **.**
> gen <- c(„Bărbat", „Bărbat", „Femeie")
> data1 <- data.frame(număr subiect, nume subiect, vârstă, sex)
> date1

	numărul	**nume subiect**	**vârstă**	**gen**
1	**1**	**Petru**	**26**	**Bărbat**
2	**2**	**Paul**	**27**	**Bărbat**
3	**3**	**Maria**	**28**	**Femeie**

>
> *# Introducerea unei funcții simple*
> odds.ratio <- function(aa, bb, cc, dd){ aa*dd / (bb*cc)}
> odds.ratio(30, 174, 21, 184) *# Date din tabelul 3.3* **.**
[1] 1,510673

INTRODUCEREA ȘI ANALIZA DATELOR FOLOSIND FUNCȚIA SCAN() (TEETOR, 2011)

Funcția R scan() este preluată din baza pachetului CRAN . Această funcție, care citește datele într-un vector sau listă din consolă sau fișier, are următoarea formă de utilizare:

```
scan(fișier = "", ce = dublu(), nmax = -1, n = -1, sep = "",
     citat = if(identic(sept, "\n")) "" else ""\"", dec = ".",
     skip = 0, nlinii = 0, na.strings = "NA",
     flush = FALSE, umplere = FALSE, strip.white = FALSE,
     silent = FALSE, blank.lines.skip = TRUE, multi.line = TRUE,
     comment.char = "", allowEscapes = FALSE, fileEncoding = "", encoding =
     "necunoscut", text)
```

Argument:

ce	Tipul a ceea ce oferă tipul de date de citit. Tipurile acceptate sunt logice, întregi, numerice, complexe, caractere, brute și listă . Dacă ceea ce este o listă, se presupune că liniile fișierului de date sunt înregistrări, fiecare conținând elemente de lungime (ce) („câmpuri"), iar componentele listei ar trebui să aibă elemente care sunt unul dintre

Argumentul what descrie jetoanele pe care scan() ar trebui să se aștepte în fișierul de intrare.

Pentru o descriere detaliată a acestei funcții, executați:

> ?scanare

Metodologia aplicării scan() este similară cu cea pentru c() , așa cum este descris în subsecțiunea precedentă, cu excepția faptului că nu contează că numerele sunt introduse pe linii diferite. Rezultatul va fi tot un vector.

> Utilizați scan() când accesați date dintr-un fișier care are o structură neregulată sau

complexă.

- Utilizați scan() pentru a citi jetoane individuale și utilizați argumentul ce să descrie fluxul de jetoane din fișier.
- scan() convertește jetoanele în date și apoi adună datele în înregistrări.
- Utilizați scan() împreună cu readLines(), mai ales când încercați să citiți un format de fișier neortodox. Împreună, aceste două funcții vor duce probabil la - procesarea cu succes a liniilor individuale și a simbolurilor fișierului.

Funcția readLines() citește linii dintr-un fișier și le returnează la o listă de șiruri de caractere:

```
> linii <- readLines("input.text")
```

Se poate limita numărul de linii de citit, pe trecere, utilizând parametrul n , care oferă numărul maxim de linii de citit:

```
> linii <- readLines("input.text, n=5) # Citiți 5 linii și opriți
```

Funcția scan() citește câte un token și îl gestionează conform instrucțiunilor. De exemplu, să presupunem că fișierul care trebuie scanat și citit conține tripleți de date (cum ar fi datele și maximele și minimele zilnice corespunzătoare ale piețelor financiare):

15-Oct-1987 2439.78 2345.63 16-Oct-1987 2396.21 2207.73
19-Oct-1987 2164.16 1677.55 20-Oct-1987 2067.47 1616.23
21-Oct-1987 2087.07 1951.76

Utilizați o listă pentru a informa scan() că ar trebui să se aștepte la o secvență repetată de 3 jetoane: > tripleți <- scan("triples.txt, what=list(character(0), numeric(0), numeric(0))))

)

Dați nume elementelor listei, iar scan() va atribui acele nume datelor: > tripleți <- scan(" triples.txt, + what=list(data=character(0), high=numeric(0), low) =numeric(0)))

Aici, se citește cinci înregistrări:

> triple # *Ieşiri:*
$date
[1] „15-oct-1987" „15-oct-1987" „19-oct-1987" „20-oct-1987" „21-oct-1987" $ mare
[2] 2439-78 2396,21 2164,16 2067,47 2081,07 $mic
[3] 2345,63 2207,73 1677,55 1616,21 1951,76

INTRODUCEREA ȘI ANALIZA DATELOR FOLOSIND SURSA FUNCȚIEI() (ARAGON , 2011; TEETOR, 2011 ; VENABLES ET AL., 2004)

Funcția R source() este de asemenea preluată din baza pachetului CRAN . Această funcție, care citește datele într-un vector sau listă din consolă sau fișier, are următoarea formă de utilizare:

source() determină R să accepte intrarea sa din fișierul sau adresa URL sau conexiune numită. Intrarea este citită și analizată din acel fișier până când se ajunge la sfârșitul fișierului; apoi expresiile analizate sunt evaluate secvenţial în mediul ales:

sursă (fișier, local = FALSE, echo = pronunțat, print.eval = ecou,
verbose = getOption ("verbos"),
prompt.echo = getOption(„prompt"), max.deparse.length = 150, chdir = FALSE, encoding = getOption(„encoding"), continue.echo = getOption(„continuare"), skip.echo = 0, păstrați .source = getOption("keep.source"))

Comenzile care sunt stocate într-un fișier extern, cum ar fi **commands.R** în directorul de lucru „work", pot fi executate într-un mediu R cu comanda:

> sursa(„comanda.R")

Funcția source() îi indică lui R să citească textul și să execute conținutul acestuia. Astfel, atunci când aveți o bucată lungă sau folosită frecvent de cod R , o puteți captura într-un fișier text. Acest lucru vă permite să rulați din nou codul fără a fi nevoie să-l introduceți din nou și să utilizați funcția source() pentru a citi și a executa codul.

De exemplu: Să presupunem că fișierul **howdy.R** conține salutul familiar:

Imprimare(„Bună, prietenul meu!")

Prin aprovizionarea fișierului, puteți executa conținutul fișierului, ca în următorul segment de cod R :

> sursa("howdy.R")

[1] "Bună, prietenul meu!"

Setarea echo-TRUE va ecou aceleași linii de script înainte de a fi executate, cu promptul R afișat înainte de fiecare linie:

```
> sursa("howdy.R", echo=TRUE)
> Imprimare(„Bună, prietenul meu!")
[1] "Bună, prietenul meu!"
```

INTRODUCEREA ȘI ANALIZA DATELOR FOLOSIND INTERFAȚA FOIA DE TABLUR ÎN R (ARAGON, 2011)

Introducerea datelor cu interfața foii de calcul a lui R utilizează următoarele funcții R din pachetul de utilitare:

data.entry(..., Modes = NULL, Names = NULL) dataentry(date, moduri)
de(..., Moduri = listă(), Nume = NULL)

Argumentele acestor funcții R sunt următoarele:

...	O listă de variabile; în prezent, aceștia ar trebui să fie vectori numerici sau de caractere sau o listă care să conțină astfel de vectori.
Moduri	Modurile de utilizat pentru variabile.
Nume	Numele care vor fi folosite pentru variabile.
Date	O listă de vectori numerici și/sau de caractere.
Moduri	O listă cu o lungime până la cea a datelor care oferă modurile (unelor dintre) variabile; list() este permisă.

Funcția data.entry() editează un obiect existent, salvând modificările la numele obiectului original. Cu toate acestea, funcția edit() editează un obiect existent, dar *nu* salvează modificările la numele obiectului original; astfel, trebuie să-l atribuiți unui nume de obiect (chiar dacă este numele original).

Pentru a introduce un vector, trebuie să inițializați un vector și apoi să folosiți datele funcției . intrare(). De exemplu:

Începeți prin a intra în mediul R și setați:

```
> x <- c(2, 4, 6, 8, 10) # X este definit inițial ca o matrice de cinci elemente .
> x # Doar verific pentru a fi sigur.
> 1] 2 4 6 8 10 # x este într-adevăr setat să fie o matrice de cinci elemente.
>
> data.entry(x) # Accesarea editorului de date:
> # Se deschide fereastra Data Editor . Privind prima coloană :
> # acum se numește „x", cu primele cinci rânduri (toate în prima coloană) completate
> #, respectiv, prin numerele 2,4, 6, 8,10
> # Acum se poate edita acest set de date, de exemplu, schimbând toate intrările la 2, apoi
> # închiderea ferestrei Editor de date și revenirea la fereastra consolei R.
> x
[1] 2 2 2 2 2 # x este într-adevăr schimbat.
■ # Astfel, se pot modifica intrările pentru x prin intermediul Editorului de date și se pot salva modificările.
```

Când utilizați funcțiile data.entry(x) și edit() pentru introducerea datelor, există o serie de limitări (Aragon, 2011) :

- Matricele și listele netabulare nu pot fi introduse folosind un editor de foi de calcul.
- Când folosiți funcția edit() pentru a crea un nou cadru de date, trebuie să îi atribuiți un nume de obiect pentru a salva cadrul de date.
- Această abordare nu este o metodă preferată pentru introducerea datelor, deoarece adesea se preferă să aibă datele originale într-un editor de text sau să fie disponibile pentru a fi citite dintr-un fișier de date.

EDITAREA UNUI CADRU DE DATE ÎN R (ADLER, 2010)

Pentru a edita un cadru de date, se poate folosi funcția edit() . Aceasta apelează un editor de foi de calcul cu o coloană pentru fiecare variabilă din cadrul de date. În foaia de calcul a editorului, se poate direcționa mouse-ul/cursorul și începe editarea celulelor existente tastând noile date în locul datelor vechi. De asemenea, se poate schimba tipul de variabilă de la real (numeric) la caracter (factor) făcând clic pe anteturile coloanei. Numele variabilelor pot fi, de asemenea, modificate. Când Editorul de date este închis, *noul cadru de date editat este atribuit noului nume dat, iar cadrul de date original este lăsat neschimbat.*

De exemplu, dacă setul de date cancer , în pachetul supraviețuire , urmează să fie editat, se poate folosi:

```
> date (cancer)
> cancer1 <- edit(cancer)
```

Acest lucru este ilustrat în exemplul următor. Mai mult, pentru a introduce date într-un cadru de date gol, se poate folosi:

```
> date noi <- data.frame()
> remediere (date noi)
```

> **Exemplu:** în cadrul de date cancer , în pachetul CRAN supraviețuire , modificați valoarea meal.cal a primului rând de la 1175 la 1176.

Soluție: Următorul segment de cod R este utilizat pentru a îndeplini sarcina de editare necesară:

```
> install.packages("survival") # Instalarea pachetului survival
> bibliotecă (supraviețuire) # Aducerea fișierelor de supraviețuire
> ls("pachet:survival") # Listarea tuturor fișierelor din supraviețuire , menționând că
>                       cancerul cadrul de date este printre ele.
```

[1] „aareg”	**"aml"**	**"attrassign"**
[4] „basehaz”	**"vezica urinara"**	**"vezica 1"**
[7] „vezica urinară2”	**"cancer"**	**"cch"**
[10] „cgd”	**"clogit"**	**"cluster"**
[13] „colon”	**„cox.zph”**	**"coxph"**
[16] „coxph.control”	**"coxph.detail"**	**"coxph.fit"**
[19] „dsurvreg”	**"format.Surv"**	**"fragilitate"**
[22] „frailty.gamma”	**"fragilitate.gaussian"**	**"fragilitate.t"**
[25] „inima”	**„este.na.coxph.penalt**	**"este.na.evaluare"**
[28] „is.na.Surv”	**"este.evaluabil"**	**"este.Surv"**
[31] „jasa”	**"jasa1"**	**"rinichi"**
[34] „labels.survreg”	**"leucemie"**	**"Logan"**
[37] „plămân”	**"match.ratetable"**	**"mgus"**
[40] „mgus1”	**"mgus2"**	**"nwtco"**
[43] „ovarian”	**"pbc"**	**"pbcseq"**
[46] „pspline”	**"psurvreg"**	**"pears"**
[49] „qsurvreg”	**"tabel de evaluare"**	**"evaluareData"**
[52] „șobolani”	**"cresta"**	**"stanford2"**
[55] „strat”	**„Surv”**	**"survConcordance**
[58] „survdiff”	**"survexp"**	**"survexp.mn"**
[61] „survexp.us”	**"survexp.usr"**	**"survfit"**
[64] „survfitcoxph.fit”	**"survobrien"**	**"survreg"**
[67] „survreg.control”	**"survreg.distributions"**	**"survreg.fit"**
[70] „survregDtest”	**"survSplit"**	**"tcut"**
[73] „tobin”	**"tt"**	**"descurca.special**
[76] „veteran”		

> data(cancer) *# Apelarea în cadrul de date* cancer
> cancer *# Verificarea cadrului de date (se uită numai la primele 5 rânduri)*

	inst	timp	starea	vârst	sex	ph.ecog	ph.karno	pat.karno	masa.cal	pierde
1	3	306	2	74	1	1	90	100	1175	N /
2	3	455	2	68	1	0	90	90	1225	15
3	3	1010	1	56	1	0	90	90	N /	15
4	5	210	2	57	1	1	90	60	1150	11
5	1	883	2	60	1	0	100	90	N /	0

> cancer1 <- edit(cancer) *# Editare* cancer *și redenumirea acestuia* cancer1 > *# O foaie de calcul cu setul de date* cancerul *se deschide. În foaia de calcul,* > *# modificați manual prima intrare* meal.cal de la *„1175" la „1176"* **:**

	inst	timp	starea	vârst	sex	ph.ecog	ph.karno	pat.karno	masa.cal	pierde
1	3	306	2	74	1	1	90	100	1176	N /
2	3	455	2	68	1	0	90	90	1225	15
3	3	1010	1	56	1	0	90	90	N /	15
4	5	210	2	57	1	1	90	60	1150	11
5	1	883	2	60	1	0	100	90	N /	0

> *# Închide foaia de calcul. Reveniți la mediul* R *și verificați* > *# cadrul de date nou editat* cancer1 **:**

> cancer1

	inst	timp	starea	vârst	sex	ph.ecog	ph.karno	pat.karno	masa.cal	pierde
1	3	306	2	74	1	1	90	100	1176	N / A
2	3	455	2	68	1	0	90	90	1225	15
3	3	1010	1	56	1	0	90	90	N / A	15
4	5	210	2	57	1	1	90	60	1150	11
5	1	883	2	60	1	0	100	90	N / A	0

> # cancer1 *este setul de date editat, după cum este necesar. Cadrul de date original*
> # cancer *rămâne neschimbat.*

Function list() și Making of data.frame() în R (Dalgaard, 2002; Teetor, 2011; Venables și colab., 2004)

de FUNCȚII ()

O listă în R constă dintr-o colecție ordonată de obiecte - **componentele sale** - care pot fi de orice mod sau tip. De exemplu, o listă poate consta dintr-o matrice, un vector numeric , un vector complex, o valoare logică, o matrice de caractere, o funcție și așa mai departe. Următorul exemplu arată o modalitate simplă de a crea o listă.

> **Exemplu:** este la fel de ușor ca 1,2, 3!

```
> x <- 1
> y <- 2
> z <- 3
> list1 <- list(x, y, z) # Formarea unei liste simple
> list1 # Ieșire:
[[1]]
[1] 1
[[2]]
[2] 2
[[3]]
[3] 3
```

Componentele sunt întotdeauna numerotate și pot fi denumite ca atare. Astfel, dacă my.special.list este numele unei liste cu patru componente, acestea pot fi denumite individual ca my.special.list[[1]], my.special.list[[2]], my.special. .list[[3]] și .lista.specială[[4]] a mea.

Dacă unul definește my.special.list după cum urmează:

```
> lista mea.specială <- list(name="John", wife="Mary",
+ numărul.de.copii=3, vârsta.copii=c(2, 4, 6))
```

apoi

```
> my.special.list[[1]] # Ieșire:
[1] "Ioan"
> lista.mea.specială[[2]]
[1] "Maria"
> lista.mea.specială[[3]]
```

> **1] 3**
> lista.mea.specială[[4]]
[1] 2 4 6

Numărul de componente dintr-o listă: numărul de componente (de nivel superior) dintr-o listă poate fi găsit prin funcția length(). Astfel:

> lungime (lista mea.specială)
> **1] 4**

Adică lista my.special.list are patru componente.

Pentru a combina un set de obiecte într-o colecție compusă mai mare pentru o procesare mai eficientă, funcția list poate fi utilizată pentru a construi o listă din componentele sale. Ca exemplu, luați în considerare

> cote <- c(1, 3, 5, 7, 9, 11,13,15,17,19)
> pare <- c(2, 4, 6, 8, 10, 12, 14, 16, 18, 20)
> lista mea <- listă(înainte=cote, după=pari)
> lista mea
$inainte
[1] 1 3 5 7 9 11 13 15 17 19
$dupa
[2] 2 4 6 8 10 12 14 16 18 20
> lista mea$inainte
[3] 1 3 5 7 9 11 13 15 17 19
> mylist$după
[1] 2 4 6 8 10 12 14 16 18 20

Componentele unei liste: Componentele unei liste pot fi numite. Într-un astfel de caz, componenta poate fi menționată fie:

1. Dând numele componentei ca șir de caractere în locul numărului dintre paranteze pătrate duble sau
2. Dând o expresie de forma > nume$nume_componentă pentru același obiect

> **Exemplul 2:** Concatenarea listelor

Luați oricare trei, mai mult sau mai puțin, liste:

> list.A <- c(„The", „rapid", „maro")
> lista.A
> **1] „Maro" „rapid"**
> listă.B <- c(„vulpe", „sări", „peste")
> lista.B
> **1] „vulpe" „sare" „peste"**
> listă.C <- c(„cel", „leneș", „câine")
> lista.C
> **1] „câinele" „leneș"**
> lista.ABC <- c(lista.A, lista.B, lista.C) *# Concatenarea celor trei liste*
> list.ABC *# Ieşire:*
[1] „Vulpea" „rapid" „maro" „sare" „peste" „câinele" „leneș" >

CONSTRUCȚIA CADRELOR DE DATE FOLOSIND FUNCȚIA data.frame()

Un *cadru de date* este o listă cu clasa „ data.frame ”. Restricțiile privind listele care pot fi transformate în cadre de date sunt următoarele:

- Componentele din listă trebuie să fie vectori (inclusiv numerice, logice sau caractere), matrice numerice, liste, factori sau alte cadre de date.
- Matricele, listele și cadrele de date furnizează noului cadru de date atâtea variabile câte coloane, elemente și, respectiv, variabile are acel cadru.
- factorii numerici și logici sunt incluși ca atare. Vectorii de caractere se limitează la a fi factori, ale căror niveluri sunt valorile unice care apar în vector.
- Structurile vectoriale care apar ca variabile ale cadrului de date ar trebui să aibă toate aceeași dimensiune de rând.

Pentru cele mai multe scopuri, un cadru de date poate fi considerat o matrice cu coloane posibil de diferite moduri și atribute. Cu toate acestea, un cadru de date poate fi afișat sub formă de matrice, cu rândurile și coloanele sale extrase folosind convențiile de indexare a matricei.

Funcția data.frame() , din baza pachetului R , creează cadre de date care sunt colecții strâns cuplate de variabile care partajează multe dintre proprietățile matricelor și ale listelor. Această funcție este utilizată ca structură de date fundamentală de majoritatea programelor de modelare ale lui R. Este definit de

```
data.frame(..., row.names = NULL, check.rows = FALSE,
                  check.names = TRUE,
                  stringsAsFactors = default.stringsAsFactors())
```

Multe reguli utilizate pentru matrice sunt aplicabile și cadrelor de date. De exemplu, structura principală a unui cadru de date constă din coloane (sau variabile) și rânduri (sau înregistrări). Regulile pentru subscriere, legarea coloanelor sau rândurilor și selectarea unui subset în matrice se aplică cadrelor de date.

Cu toate acestea, cadrele de date sunt mai complicate decât matricele:

- Toate coloanele dintr-o matrice sunt forțate să fie caracter dacă doar o celulă este un caracter.

  În schimb, un cadru de date poate avea diferite clase de coloane. De exemplu, un cadru de date poate consta dintr-o coloană „Patient.ID” (care este numerică); și o coloană „nume” (care este caracter).
- Un cadru de date poate avea și atribute suplimentare. De exemplu, fiecare variabilă poate avea descrieri lungi ale variabilelor.
- Un factor dintr-un cadru de date are adesea „niveluri” sau etichete de valoare.

Aceste atribute pot fi transferate din setul de date original în alte formate, cum ar fi SAS, SPSS sau Stata. Ele pot fi create și în R în timpul analizei.

OBȚINEREA UNUI CADRU DE DATE DIN UN FIȘIER TEXT

Obținerea unui cadru de date dintr-un fișier text a fost discutată mai devreme în această secțiune și au fost date exemple. Acele subsecțiuni s-au ocupat de crearea cadrelor de date prin citirea datelor create din programe din afara lui R , cum ar fi Excel pe platforma Windows. De asemenea, am discutat și am dat exemple despre obținerea unui cadru de date prin introducerea datelor direct în R folosind funcția data.entry() . După cum s-a menționat acolo, șansa de eroare umană este mare atunci când se efectuează introducerea datelor în foaie de calcul sau în modul text. Prin urmare, programul software *EpiData* ,

care este special conceput pentru introducerea datelor, este mai potrivit.

EpiData are facilități pentru stabilirea constrângerilor utile, cum ar fi verificări ale intervalului, salturi automate și etichetarea variabilelor și a valorilor (codurilor) pentru fiecare variabilă. Se poate face un transfer direct între *EpiData* și R (folosind „read.epiinfo"), dar se recomandă să exportați datele din *EpiData* (folosind procedura de export din acel software) în format Stata și să utilizați funcția read.dta pentru a citi set de date în R. Exportarea datelor în format Stata menține multe dintre atributele variabilelor, cum ar fi etichetele și descrierile variabilelor (Virasakdi, 2011) .

Întrebări de revizuire pentru Secțiunea 3.4

1. Pentru a utiliza R în analiza datelor în BIOS, datele de procesat trebuie mai întâi introduse în mediul R. Discutați șapte moduri de introducere a datelor, dând exemple.
2. Cum poate fi folosită funcția list() pentru a introduce date în mediul R ? Dați un exemplu.
3. Utilizați funcția data.frame() pentru a introduce date în mediul R , dând un exemplu.
4. Utilizați următoarele funcții pentru a introduce date în mediul R , dând un exemplu pentru fiecare: c() , matrix() și array() .
5. Utilizați funcția source() pentru a introduce date în mediul R , dând un exemplu.
6. Care sunt limitările atunci când utilizați funcțiile data.entry(x) și edit() pentru introducerea datelor?
7. Arătați că funcția list() poate fi folosită pentru a combina mai multe componente pentru a forma o nouă listă, dând un exemplu.
8. Scrieți un segment de cod în R pentru a extrage numele unei componente stocate într-o altă variabilă, dând un exemplu.
9. Configurați un exemplu în care utilizați funcția de concatenare c() cu argumente de listă date și obțineți o listă ale cărei componente sunt cele ale listei de argumente unite secvențial, sub următoarea formă:
   > lista.ABC <- c(lista.A, lista.B, lista.C)
10. Căutați software-ul *EpiData* de pe site-ul său web (www.epidata.dk) și sugerați o metodă eficientă de introducere a datelor în R .

Exerciții pentru Secțiunea 3.4 (Dalgaard, 2002; Everitt & Hothorn , 2006 ; Virasakdi, 2011)

1. Datele despre cancerul vezicii urinare în HSAUR2 (Everitt & Hothorn, 2006) .
   Datele au fost preluate de la 31 de bărbați care au fost tratați pentru cancer superficial de vezică urinară. Datele înregistrează numărul de tumori recurente în timpul unui anumit

   timp după îndepărtarea tumorii primare, împreună cu dimensiunea tumorii inițiale. Să aruncăm o privire în profunzime asupra setului de date.

   (a) Utilizați următoarele segmente de cod pentru a introduce setul de date pentru a-i examina conținutul:

   > date("cancer de vezică", pachet = "HSAUR2")

   (b) Ieșiți întregul cadru de date folosind segmentul de cod:

> cancerul vezicii urinare

(c) Sortați cadrul de date după componente folosind segmentul de cod:

> data1 <- c(~ număr + mărimea tumorii, date = cancer de vezică)
> date1

(d) Folosind următorul segment de cod, puteți ajunge și la pachetul HSAUR2 , care conține setul de date bladdercancer .

> install.packages("HSAUR2")

Vi se va cere să selectați un site CRAN pentru susținerea muncii dvs.:

Vă rugăm să selectați o oglindă CRAN pentru a fi utilizată în această sesiune --

Dacă nu aveți o preferință anume, selectați site- ul SUA (CA1) , care este Universitatea din California din Berkeley:

> date („cancer de vezică")

2. Iată un alt cadru de date:

(a) Utilizați următoarele segmente de cod pentru a introduce setul de date pentru a-i examina conținutul:

> date("HELPrct", pachet = "mozaic")
> date

(b) Utilizați următorul segment de cod pentru a extrage componenta din setul de date:

> data1 <- c(data = HELPrct)
> date1

3. Iată încă un cadru de date de interes epidemiologic: cadrul de date USmelanoma , care poate fi descărcat și din pachetul R HSAUR2 .

În acest caz, se poate examina rata anuală a mortalității de la un melanom malign în funcție de statul american, precum și de latitudinea centrelor lor geografice. Aceste date epidemiologice au fost colectate de la populația de bărbați albi din Statele Unite în perioada 1950-1969. Studiul a fost de interes pentru că s-a crezut că

- Persoanele cu culoarea deschisă a pielii sunt mai susceptibile la dezvoltarea melanomului malign.
- Latitudinea geografică a locației este legată de cantitatea de expunere la soare pe unitatea de timp în medie.

(a) Utilizați următoarele segmente de cod pentru a descărca setul de date din HSAUR2 :

> date ("USmelanoma", pachet = "HSAUR2")

(b) Utilizați următorul segment de cod pentru a extrage setul de date USmelanoma :

> date
[1] „USmelanom"
> USmelanomul

(c) Sortați cadrul de date după componente folosind următorul segment de cod:

> data1 <- c(~ mortalitate + latitudine, date = USmelanom)
> date1

4. Iată câteva cadre de date din studiul geneticii (date despre mărimea genomului de la www.ornl.gov; Seefeld & Linder, 2005 *)* .

(a) Utilizați următoarele segmente de cod pentru a descărca vectorii care constituie

setul de date și pentru a crea obiectul cadru de date:

```
> organism<-c("Uman","Soarece","Musca de fructe",
+                  „Viermi rotunzi", „Drojdie")
> genomeSizeBP<-
+    c(3000000000,3000000000,135600000,97000000,12100000)
> estGeneCount<-c(30000,30000,13061,19099,6034)
```

(b) Dacă aveți trei vectori de lungime egală, îi puteți uni într-un cadru de date folosind funcția data.frame() cu vectorii ca argumente ale acestei funcție. Rețineți că formatul este „nume coloană"="vector de adăugat" și este utilizat operatorul egal (nu alocare <-) . Aici, numiți coloane mai degrabă decât creați variabile noi: numele variabilelor sunt folosite ca nume de coloane, dar puteți redenumi coloanele cu alte nume decât numele variabilelor: > comparativeGenomeSize<- + data.

```
> comparativeGenomeSize<-
+ data.frame(organism=organism,genomeSizeBP=genomeSizeBP,
+ estGeneCount=estGeneCount)
> comparativeGenomeSize
```

(c) Sortați cadrul de date după componente folosind următorul segment de cod: > data1 <- c(~ organism, data = comparativeGenomeSize) > data1

5. *Utilizarea editorului de date: de* la ISwR (*Introductory* S *tatistics with* R ; Dalgaard , 2002)

(a) Aduceți pachetul ISwR în mediul R computer-desktop folosind următorul segment de cod:

```
> install.packages ("ISwR")
> # Pentru --- Vă rugăm să selectați o oglindă CRAN pentru a fi utilizată
```

în această sesiune --- *# selectați CA1 (b) Editați* calitatea aerului cadrului de date folosind funcția edit() în următorul segment de cod:

```
> date (calitatea aerului)
> aq <- edita (calitate aer)
```

Aceasta afișează un editor asemănător unei foi de calcul cu o coloană pentru fiecare variabilă din cadrul de date.

(c) Odată ajuns în editor, deplasați cursorul cu mouse-ul sau touch pad-ul și editați celulele introducând date noi sau modificând datele existente. Apoi închideți editorul și introduceți:

```
> aq
```

Primele două rânduri ale fișierului aq vor arăta astfel:

	Ozon	Solar.R	Vânt	Temp	Lună	Zi
1	**41**	**190**	**7.4**	**67**	**5**	**10**
2	**36**	**118**	**8.0**	**72**	**5**	**2**

(d) Reveniți la Editorul de date introducând:

```
> aq <- edit (calitate aer)
```

Odată revenit în editor, modificați datele. De exemplu, modificați primul rând al variabilei „Ozon" de la 41 la 42. Apoi închideți din nou editorul și introduceți:

> aq

Primele două rânduri ale fișierului aq vor arăta astfel:

	Ozon	Solar.R	Vânt	Temp	Lună	Zi
1	42	190	7.4	67	5	10
2	36	118	8.0	72	5	2

Valoarea variabilei „Ozon" de pe primul rând a fost modificată de la 41 la 42. Procesul de editare a datelor a fost un succes?

Observație: Când Editorul de date este închis, cadrul de date editat este atribuit lui aq și calitatea aerului fișierului original este lăsată neschimbată.

3.5 ANALIZA DATELOR UNIVARIATĂ, BIVARIATĂ ȘI MULTIVARIATĂ

Un set de date **univariat** are o singură variabilă: { x }, cum ar fi { *numele pacientului* }.

Un set de date **bivariat** are două variabile: { x_1 , x_2 } sau { x , y }, cum ar fi { *numele pacientului* , *sexul* }.

Un set de date **multivariat** are mai mult de două sau mai multe variabile: { x_1 , x_2 , x_3 , …, x_n }, cum ar fi { *numele pacientului* , *sexul* , *vârsta* , *diagnosticul* , *tratamentul* , …}.

Analiza univariată a datelor

Ca exemplu, introduceți următoarele segmente de cod:

```
> x <- rex(100); x # Se produc 100 aleatoriu distribuit exponențial
>                  # numere:
  [1] 0,39136880 0,66948212 1,48543076 0,34692128 0,71533079 0,12897216
  [7] 1,08455419 0,07858231 1.01995665 0,81232737 0,78253619 4,27512555
 [13] 2.11839466 0,47024886 0,62351482 1,02834522 2.17253419 0,37622879
 [19] 0,16456926 1,81590741 0,16007371 0,95078524 1.26048607 5,92621325
 [25] 0,21727112 0,07086311 0,83858727 1,01375231 1,49042968 0,53331210
 [31] 0,21069467 0,37559212 0,10733795 2,84094906 0,17899040 1,34612473
 [37] 0,00290699 1,77078060 1,79505318 0,09763821 1,96568170 0,15911043
 [43] 4,36726420 0,33652419 0,01196883 0,35657882 0,72797670 0,91958975
 [49] 0,68777857 0,29100399 0,22553560 1,56909742 0,20617517 0,37169621
 [55] 0,53173534 0,26034316 0,21965356 2,94355695 1,88392667 1.13933083
 [61] 0,31663107 0,23899975 0,01544856 1,30674088 0,53674598 1.72018758
 [67] 0,31035278 0,81074737 0,09104104 1,52426229 1.35520172 0,27969075
 [73] 1.36320488 0,56317216 0,85022837 0,49031656 0,17158651 0,31015165
 [79] 2.07315953 1,29566872 1,28955269 0,33487343 0,20902716 2,84732652
 [85] 0,58873236 1,54868210 2,93994181 0,46520037 0,73687959 0,50062507
 [91] 0,20275282 0,49697531 0,58578119 0,49747575 1,53430435 4,56340237
 [97] 0,90547787 0,72972219 2.60686316 0,33908320
```

Notă: Funcția rex() este definită după cum urmează: rex(n, rata = 1)

cu argumente:

x	Vector
n	Numărul de observații. Dacă lungimea(n) > 1, lungimea este considerată numărul necesar.

Distribuția exponențială cu rata *l* are următoarea densitate: $f(x) = l\, e^{-l x}$, pentru $x \geq 0$. Dacă rata *l* nu este specificată, aceasta presupune valoarea implicită de 1.

Observație: Funcția reexp() este una dintre funcțiile din R sub Exponențial în statisticile pachetului CRAN .

Pentru a efectua o analiză biostatistică a acestui set de date univariate, apelați funcția univax(), din pachetul epibasix, folosind următoarele segmente de cod: > library(epibasix)

> univar(x) # *Ieșire:*

Rezumat univariat

Dimensiunea eșantionului: 100

Media eșantionului: 1,005

Mediana eșantionului: 0,646

Abatere standard eșantion: 1,067 >

au fost calculate media, mediana și abaterea standard .

Pentru analiza datelor seturi de date univariate, poate fi utilizat pachetul R epibasix . Acest pachet CRAN (nd) acoperă multe funcții epidemiologice elementare pentru biostatistică și epidemiologie. Conține instrumente elementare pentru analiza problemelor epidemiologice comune , variind de la estimarea dimensiunii eșantionului , până la analiza tabelului de contingență 2 × 2 și măsuri de bază de acord (kappa, sensibilitate/specificitate).

De asemenea, au fost scrise declarații adecvate tipărite și rezumate pentru a facilita interpretarea ori de câte ori este posibil. Această lucrare este adecvată pentru cursurile absolvente de biostatistică/epidemiologie. Acest pachet este un lucru în curs.

Pentru a începe, introduceți mediul R și utilizați următorul segment de cod:

> install.packages("epibasix")

Instalarea pachetelor în „C:/Users/bertchan/Documents/R/win-library/2.14” (deoarece „lib” nu este specificat)

> **-- Vă rugăm să selectați o oglindă CRAN pentru a fi utilizată în această sesiune ---**

> # *Selectați CA1*

URL-ul de încercare „ http://cran.cnr.Berkeley.edu/bin/windows/contrib/2.14/epibasix_1.1.zip” Tipul de conținut „application/zip” lungime 57888 octeți (56 Kb)

URL deschis

descărcat 56 Kb

pachetul „epibasix” a fost despachetat cu succes și sumele MD5 au fost verificate

Pachetele descărcate sunt în C:\Utilizatori\bertchan\AppData\Local\Temp\RtmpMFOrEn\downloaded_ pachete

Cu epibasix încărcat în mediul R , urmați acești pași pentru a afla mai multe despre acest pachet:

1. Accesați site-ul web CRAN (http://cran.r-project.org).
2. Selectați (un singur clic) „Pachete" în coloana din stânga.
3. Pe pagina „Pachete", selectați E (pentru epibasix).

Pachete CRAN disponibile după nume
ABCDEFGHIJKLMNOPQRSTU VWXYZ

4. Derulați în jos lista pachetelor ale căror nume încep cu „E" sau „e" și selectați:

epibasix	Funcții epidemiologice elementare pentru un curs de absolvent de epidemiologie/biostatistică

5. Când se deschide pagina epibasix , selectați: Manual de referință: ***epibasix.pdf***
6. Informațiile sunt acum afișate după cum urmează:

Pachet	**"epibasix"**
	2 ianuarie 2012
Versiune	**1.1**
Data Titlu	**13-05-2009**
	Funcții epidemiologice elementare pentru un absolvent
	Curs de epidemiologie\{}Biostatistică
Autorul de	**Michael A Rotondi < mrotondi@uwo.ca> Michael A**
întreținere	**Rotondi** mrotondi@uwo.ca **(>= 2,01)**
depinde de R	

Pentru un alt exemplu, luați în considerare aceeași analiză asupra primelor 100 de numere naturale , folosind următoarele segmente de cod R :

```
> x <-1:100; x # Luați în considerare și apoi scoateți primele 100 de numere naturale:
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18
[19] 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
[37] 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54
[55] 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72
[73] 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90
[91] 91 92 93 94 95 96 97 98 99 100
  > # Tabele ANOVA: Rezumat în următoarele tabele, ANOVA este utilizat pentru
  > # două scopuri diferite:
  > bibliotecă (epibasix)
  > univar(x) # Efectuarea unei analize de date univariate pe vectorul x, și
  > # ieșire:
Rezumat univariat
Dimensiunea eșantionului: 100

Medie eșantion: 50,5
Mediana eșantionului: 50,5
Abatere standard eșantion: 29,011
```

Analiza datelor bivariate și multivariate (Daniel, 2005)

Când există două variabile, (X , Y), trebuie să luăm în considerare două cazuri:

Cazul I : În modelul clasic de regresie, numai Y (numit *variabilă dependentă*) trebuie să fie aleatoriu. X este definit ca o variabilă fixă, non-aleatoare, și se numește *variabilă independentă* . Conform acestui model, observațiile sunt obținute prin preselectarea valorilor lui X și determinarea valorii corespunzătoare a lui Y.

Cazul II: Dacă atât X cât și Y sunt variabile aleatoare, se folosește **modelul de corelație .** În cadrul acestui model, observațiile eșantionului sunt obținute prin selectarea unui eșantion aleatoriu al unităților de asociere, cum ar fi persoane, caracteristici (vârstă, sex, locații, momente de timp, evenimente/acțiuni specifice/…,) sau elemente pe care cele două măsurătorile se bazează; și prin înregistrarea unei măsurători a *lui X* și a lui Y . În acest caz, valorile lui X nu sunt preselectate, ci mai degrabă apar la întâmplare, în funcție de unitatea de asociere selectată în eșantion.

ANALIZA REGRESIEI

Cazul I — Analiza corelației nu poate fi realizată în mod semnificativ în cadrul acestui model. **Cazul II** —Analiza de regresie poate fi efectuată sub modelul de corelație.

Corelația pentru două variabile implică o corelație între variabile și nu face distincție între ele cu privire la care este variabila dependentă sau independentă. Astfel, se poate potrivi o linie dreaptă datelor fie prin minimizarea $\sum (x_i - \underline{x})^2$, fie prin minimizarea $\sum (y_i - \underline{y})^2$. Linia de regresie ajustată va fi, în general, diferită în cele două cazuri, așa că apare o întrebare logică cu privire la ce linie să se potrivească.

Există două situații și ar trebui luate în considerare:

1. Dacă obiectivul este de a obține o măsură a puterii relației dintre cele două variabile, nu contează ce linie este potrivită; măsura calculată va fi aceeași în ambele cazuri.
2. Dacă trebuie să folosiți ecuația care descrie relația dintre cele două variabile pentru a măsura dependența uneia de cealaltă, contează *ce* linie urmează să fie ajustată. Variabila pentru care se dorește să se estimeze medii sau să se facă predicții ar trebui tratată ca variabilă dependentă. Adică, această variabilă ar trebui regresată în raport cu cealaltă variabilă.

R DISPONIBILE PENTRU ANALIZA DATELOR BIVARIE

Printre pachetele R pentru analiza datelor bivariate, unul notabil disponibil pentru calculele dimensiunii eșantionului în modelul de regresie bivariat cu interceptare aleatoare (RI) este bivarRIpower .

- **analiză a datelor bivariate :** ca exemplu, acest pachet poate fi utilizat pentru a calcula dimensiunea necesară a eșantionului pentru a obține o putere de 80% la un nivel alfa de 5% pentru ipotezele nule și alternative, astfel încât corelația dintre RI să fie 0 și, respectiv, 0,2 în șase momente de timp. Alți parametri de covarianță sunt setați după cum urmează:

- Corelația dintre reziduuri = 0
- standard : primul RI = 1,2nd RI = 2, primul rezidual = 0,5, al 2-lea rezidual = 0,75

Se poate utiliza următorul segment de cod R :

- bibliotecă (bivarRIpower)
- bivarcalcn(power=.80,powerfor='RI',timepts=6,d1 = 1,d2=2,
+ p=0,p1=.2,s1=.5,s2=.75,r=0,r1=.1) *# Ieșire:*

Parametrii de variație

Clustere = 209,2
Măsurători repetate = 6
Abateri standard
Prima interceptare aleatorie = 1
A doua interceptare aleatorie = 2
Primul termen rezidual = 0,5
Al 2-lea termen rezidual = 0,75
Corelații
RI sub H_o = 0
RI sub H_a = 0,2
Reziduu sub H_o = 0
Reziduu sub H_a = 0,1
Con obs sub H_o = 0
Con obs sub H_a = 0,1831984
Lag obs sub H_o = 0
Lag obs sub H_a = 0,1674957
Variante de corelație sub H_o

Interceptare aleatorie = 0,005096138
Reziduu = 0,0009558759
Observații concurente = 0,00358999
Observații întârziate = 0,003574277
Puterea (%) pentru corelații

Interceptare aleatorie = 80%
Reziduu = 89,9%
Observații concurente = 86,4%
Observații întârziate = 80%
>

DISTRIBUȚIE NORMALĂ BIVARIATĂ

Conform modelului de corelație, bivariatele *X* și *Y* variază împreună într-o distribuție *comună* , care, dacă această distribuție comună este normală, se numește **distribuție normală bivariată** . Din această distribuție, se pot face inferențe pe baza rezultatelor eșantionării adecvate din populație. Dacă distribuția comună este cunoscută a fi nenormală sau dacă forma este necunoscută, procedurile inferențiale sunt invalide. La eșantionarea dintr-o distribuție bivariată, următoarele ipoteze trebuie să fie valabile dacă inferențe despre populație trebuie să fie valide:

- Pentru fiecare valoare a lui X, există o subpopulație distribuită normal de valori Y.
- Pentru fiecare valoare a lui Y, există o subpopulație distribuită normal de valori X.
- Distribuția comună a lui X și Y este o distribuție normală (distribuția normală bivariată).
- Subpopulația de valori Y are toate aceeași varianță.

- Subpopulația de valori X are toate aceeași varianță.

două variabile aleatoare *X* și *Y* sunt **normale în comun** dacă pot fi exprimată în formă

$$X = aU + bV \tag{3,5-1}$$

$$Y = cU + dV \tag{3,5-2}$$

unde *U* și *V* sunt variabile aleatoare normale independente.

Dacă *X* și *Y* sunt comune normale, atunci orice combinație liniară

$$Z = s_1 X + s_2 Y \tag{3,5-3}$$

are o distribuție normală. Motivul este că dacă există $X = aU + bV$ și $Y = cU + dV$ pentru unele variabile aleatoare normale independente *U* și *V* , atunci

$$Z = s_1 (aU + bV) + s_2 (cU + dV) = (ca_1 + cs_2) U + (bs_1 + ds_2) \tag{3,5-4}$$

Astfel, *Z* este suma variabilelor aleatoare normale independente $(ca_1 + cs_2) U$ și $(bs_1 + ds_2) V$ și, prin urmare, este normală.

O proprietate foarte importantă a variabilelor aleatoare comune normale este că **relația zero cor implică independență** . Dacă două variabile aleatoare *X* și *Y* sunt comune normale și sunt necorelate, atunci ele sunt independente. Această proprietate poate fi verificată folosind transformări multivariate.

Analiza multivariată a datelor (Daniel, 2005)

Două abordări similare, dar distincte, sunt utilizate pentru analiza multivariată a datelor: analiza de regresie liniară multiplă și analiza modelului de corelație multiplă.

ANALIZA REGRESIUNII LINARE MULTIPLE

Analiza de regresie liniară multiplă presupune că există o relație liniară între o variabilă Y (variabila dependentă) și n variabile independente X_1 , X_2 , X_3 , ... , X_n , care sunt numite *variabile explicative* sau predictoare din cauza felului în care sunt folosit.

Pentru regresia liniară multiplă, ecuația modelului este

$$y_j = b_0 + b_1 x_{1j} + b_2 x_{2j} + b_3 x_{3j} + \dots + b_n x_{nj} + e_j \tag{3,5-5}$$

unde y_j este o valoare tipică dintr-una dintre subpopulațiile de valori *Y* și valorile b_i sunt coeficienții de regresie.

x_{1j} , x_{2j} , x_{3j} , …, x_{nj} sunt valori particulare ale variabilelor independente X_1 , X_2 , X_3 , …, X_n , respectiv, și e_j este o variabilă aleatoare cu media 0 și varianța b^2, varianța comună a subpopulației de valori *Y*. În general, e_j se presupune a fi normal și distribuit independent.

Când ecuația (3.5-1) constă dintr-o variabilă dependentă și două variabile independente, modelul devine

$$y_j = b_0 + b_1 x_{1j} + b_2 x_{2j} + e_j \tag{3,5-6}$$

Un plan în spațiu tridimensional (3D) poate fi montat pe punctele de date. Pentru modelele care conțin mai mult de două variabile, este un **hiperplan .**

Parametrul de interes în acest model este coeficientul de **determinare multiplă** , $R^2_{y,12\dots n}$, obținut prin împărțirea sumei explicate a pătratelor la suma totală a pătratelor:

$$R^2_{y,12\ldots n} = \sum (\underline{y}_i \underline{y})^2 / \sum (\underline{y}_i \underline{y})^2 = \text{SSR/SSE} \qquad (3,5\text{-}7)$$

unde

$\sum (\underline{y}_i \underline{y})^2$ = variația explicată,
= valorile observate inițiale din valorile *Y calculate* ,
= suma abaterii pătrate a valorilor calculate față de media valorilor *Y observate*
sau
= suma pătratelor datorată regresiei (SSR)

$\sum (\underline{y}_i \underline{y})^2$ = variația inexplicabilă,
= suma abaterilor pătrate ale observațiilor originale de la valorile calculate,
= suma pătratelor despre regresie, sau
= eroarea suma pătratelor (SSE).

Variația totală este suma abaterilor pătrate ale fiecărei observații a lui *Y* față de media observațiilor:

$$\sum (y_j \underline{y})^2 = \sum (\underline{y}_i \underline{y})^2 + \sum (\underline{y}_i \underline{y})^2 \qquad (3,5\text{-}8A)$$

$$\text{SST} = \text{SSR} + \text{SSE} \qquad (3,5\text{-}8B)$$

sau

Suma totală a pătratelor = Suma explicată (regresiune) a pătratelor
+ Sumă inexplicabilă (eroare) a pătratelor (3,5-8C)

ANALIZA MODELULUI DE CORELATII MULTIPLE

Obiectivul acestei abordări este de a obține o perspectivă asupra forței relației dintre variabile.

Ipotezele care stau la baza **analizei modelului de regresie multiplă** sunt:

1. Itemii Xi sunt variabile fixe non-aleatoare, indicând că orice inferențe extrase din datele eșantionului se aplică numai setului de valori X observate, dar nu și colecțiilor mai mari de X. În cadrul acestui model de regresie, analiza corelației nu este semnificativă.
2. Pentru fiecare set de valori X_i , există o subpopulare de valori *Y*. De obicei, se presupune că aceste valori *Y* sunt distribuite în mod normal.
3. Variantele lui *Y* sunt toate egale.
4. Valorile *Y* sunt independente de diferitele seturi selectate de valori *X*.

Ecuația modelului de corelație multiplă este

$$y_j = b_0 + b_1 x_{1j} + b_2 x_{2j} + b_3 x_{3j} + \ldots + b_n x_{nj} + e_j \qquad (3,5\text{-}9)$$

unde y_j este o valoare tipică dintr-una dintre subpopulațiile de valori *Y ;* valorile b_i sunt coeficienții de regresie; x_{1j} , x_{2j} , x_{3j} , …, x_{nj} sunt valori deosebit de cunoscute ale variabilelor aleatoare X_1 , X_2 , X_3 , …, respectiv X_n ; și e_j este o variabilă aleatoare cu media 0 și varianța s^2, varianța comună a subpopulației de valori *Y*. În general, e_j se presupune a fi normal și distribuit independent.

Acest model este similar cu ecuația (3.5-5) , cu o distincție importantă: în ecuația (3.5-5) , itemii X_i sunt variabile non-aleatoare, dar în ecuația (3.5-9) , itemii X_i sunt variabile aleatoare. Adică, în modelul de corelație Ecuația (3.5-9) , există o distribuție comună a *lui Y* și X_i , care se numește **distribuție multivariată** .

Sub acest model, variabilele nu mai sunt considerate ca fiind dependente sau independente, deoarece logic sunt interschimbabile, iar orice X_i poate juca rolul lui Y.

Pentru a analiza relațiile dintre variabile, luăm în considerare **coeficientul de relație multiplă** , care este rădăcina pătrată a coeficientului de determinare multiplă . Prin urmare, valoarea eșantionului poate fi calculată luând rădăcina pătrată a ecuației (3.5-7) :

$$R^2_{y,12\ldots n} = \sqrt{R^2_{y,12\ldots n}} = \sqrt{\{\sum (\underline{y}_i \underline{y})^2 / \sum (\underline{y}_i \underline{y})^2\}} = \sqrt{(\text{SSR/SSE})} \quad (3,5\text{-}10)$$

Analiza varianței (ANOVA, nd)

În biostatistică, **AN** alysis **Of VA** riance (ANOVA) este o colecție de modele biostatistice în care varianța observată într-o anumită variabilă este împărțită în componente din diferite surse de variație . ANOVA oferă un test biostatistic dacă mediile mai multor grupuri sunt toate egale și, prin urmare, generalizează testul t la mai mult de două grupuri. Efectuarea mai multor teste *t* cu două eșantioane ar duce la o șansă crescută de a comite o eroare de tip I. Din acest motiv, ANOVA este utilă în compararea a două, trei sau mai multe medii.

După cum este rezumat în tabelele 3.4 și 3.5 , ANOVA este utilizată în două scopuri diferite:

1. Estimarea și testarea ipotezelor pentru regresia liniară simplă despre variațiile populației
2. Estimarea și testarea ipotezelor despre mediile populației

TABEL 3.4 Tabel ANOVA pentru testarea ipotezelor despre regresia liniară simplă

SURSĂ	DF	SUMA PATRATURILOR	PĂTRATE MEDIE	VALOAREA F	VALOAREA P
Model	1	$\Sigma (\hat{y}_i - \underline{y})^2 =$ SSModel	SSM = MSM	MSG/MSE = $F1_{,n-2}$	Pr(F > $F_{1, n-2}$*)
Rezidual	n - 2	Σe_i^2 = SSResidual	SSR/(n - 2) = MSE		
Total	n - 1	$\Sigma (\hat{y}_i - \underline{y})^2 =$ SSTotal	SST/(n - 1) = MST		

Reziduurile sunt adesea numite *erori* , deoarece sunt o parte a variației pe care linia nu a putut-o explica. În acest caz,

MSR = MSE = suma reziduurilor pătrate/df = s^2 = estimarea varianței dreptei de regresie a populației

SSTot/(n - 1) = MSTOT = s_y^2 = varianța totală a *y* s

F = t^2 pentru regresie liniară simplă

Cu cât este mai mare F (cu cât valoarea *p este mai mică*), cu atât linia explicată este mai mare din variația lui *y , deci cu atât este mai puțin probabil ca* H_0 să fie adevărat. Respingem o ipoteză când valoarea $p < \alpha$:

R^2 = proporția variației totale a *lui y* explicată prin linia de regresie = SSM/SST = 1 – SSResidual/SST

Statistica testului F are două DF diferite: numărătorul = $k - 1$ și denom inatorul = $N - k \rightarrow F_{k-1, N-k}$.

Notă: SSE/(N - k) = MSE = s^2 = (varianța eșantionului cumulat) = $\frac{22}{(n_1 - 1)s_1 + \ldots + (n_k - 1)s_k}(n_1)$

$$- 1) + ... + (n_k - 1)$$

$= s \hat{}^2 =$ estimare pentru varianța egală presupusă (aceasta este varianța „medie" pentru fiecare grup)
SSTot/(N - 1) = MSTOT = s^2 = varianța totală a datelor (presupunând grupuri de NO)
F ≈ varianța mediilor (între) eșantionului împărțit la varianța medie aproximativă a datelor; cu cât este mai mare F (cu cât valoarea *p este mai mică*), cu atât mediile sunt mai variate, deci cu atât este mai puțin probabil ca H_0 să fie adevărat. Este respins când valoarea $p < \alpha$.

TABEL 3.5 Tabel ANOVA pentru testarea ipotezelor despre mediile populației

SURSĂ	DF	SUMA DE PĂTRAȚELE ÎNSEAMĂ	PĂTRATE	VALOAREA F	VALOAREA P
Grup (între)	k - 1	$\Sigma n_i (\bar{x}_i - \bar{x})^2 = SSG$	$SSG/(k - 1) = MSG$	$MSG/MSE = Fk_{-1, N-k}$	$Pr(F > F_{k-1, N-k})$
Eroare (înăuntru)	N - k	$\Sigma (n_i - 1)s_i^2 = SSE$	$SSE/(N - k) = MSE$		
Total	N - 1	$\Sigma (x_{ij} - \bar{x})^2 =$ SSTot	$SSTot/(N - 1) = MST$		

Notă: N = numărul total de observații = Σn_i, unde n_i = numărul de observații pentru grupul i .

R^2 = proporție din variația totală explicată prin diferența de medii = $\frac{SSG}{SSTot}$

> **Exemplul 1 în analiza multivariată a datelor:** Studiu epidemiologic de fibroză chistică (Dalgaard, 2002)

Luați în considerare un studiu epidemiologic al funcției pulmonare la pacienții care suferă de fibroză chistică . Cadrul de date poate fi obținut de la CRAN utilizând următoarele segmente de cod R :

```
> install.packages ("ISwR")
> -- Vă rugăm să selectați o oglindă CRAN pentru a fi utilizată în această sesiune
---
> # Selectați CA1
> bibliotecă (ISwR)
> date (chistfibr)
> cystfibr # Emiterea unei matrice cu un titlu de 10 articole și 25 de cazuri :
```

	vârstă	sex	înălțime	greutate	bmp	fev1	rv	frc	tlc	pemax
1	7	0	109	13.1	68	32	258	183	137	95
2	7	1	112	12.9	65	19	449	245	134	85
3	8	0	124	14.1	64	22	441	268	147	100
4	8	1	125	16.2	67	41	234	146	124	85
5	8	0	127	21.5	93	52	202	131	104	95
6	9	0	130	17.5	68	44	308	155	118	80
7	11	1	139	30.7	89	28	305	179	119	65
8	12	1	150	28.4	69	18	369	198	103	110
9	12	0	146	25.1	67	24	312	194	128	70

10	13	1	155	31.5	68	23	413	225	136	95
11	13	0	156	39.9	89	39	206	142	95	110
12	14	1	153	42.1	90	26	253	191	121	90
13	14	0	160	45.6	93	45	174	139	108	100
14	15	1	158	51.2	93	45	158	124	90	80
15	16	1	160	35.9	66	31	302	133	101	134
16	17	1	153	34.8	70	29	204	118	120	134
17	17	0	174	44,7	70	49	187	104	103	165
18	17	1	176	60.1	92	29	188	129	130	120
19	17	0	171	42.6	69	38	172	130	103	130
20	19	1	156	37.2	72	21	216	119	81	85
21	19	0	174	54.6	86	37	184	118	101	85
22	20	0	178	64,0	86	34	225	148	135	160
23	23	0	180	73,8	97	57	171	108	98	165
24	23	0	175	51.1	71	33	224	131	113	95
25	23	0	179	71,5	95	52	225	127	101	195

```
> # Pentru a obține diagrame de dispersie pe perechi între toate variabilele
> par(mex=0,5)
> perechi (chistfibr, gap=0, cex.labels=0,9)
> # Cifra de ieșire
```

Trasarea cadrului de date: Un grafic simplu al tuturor celor 10 componente ale datelor poate fi obținut prin următorul segment de cod:

```
> plot(cystfibr) # Ieșire Figura 3.16 .
```

Funcția plot() este generică, comportându-se diferit în funcție de clasa argumentelor sale.

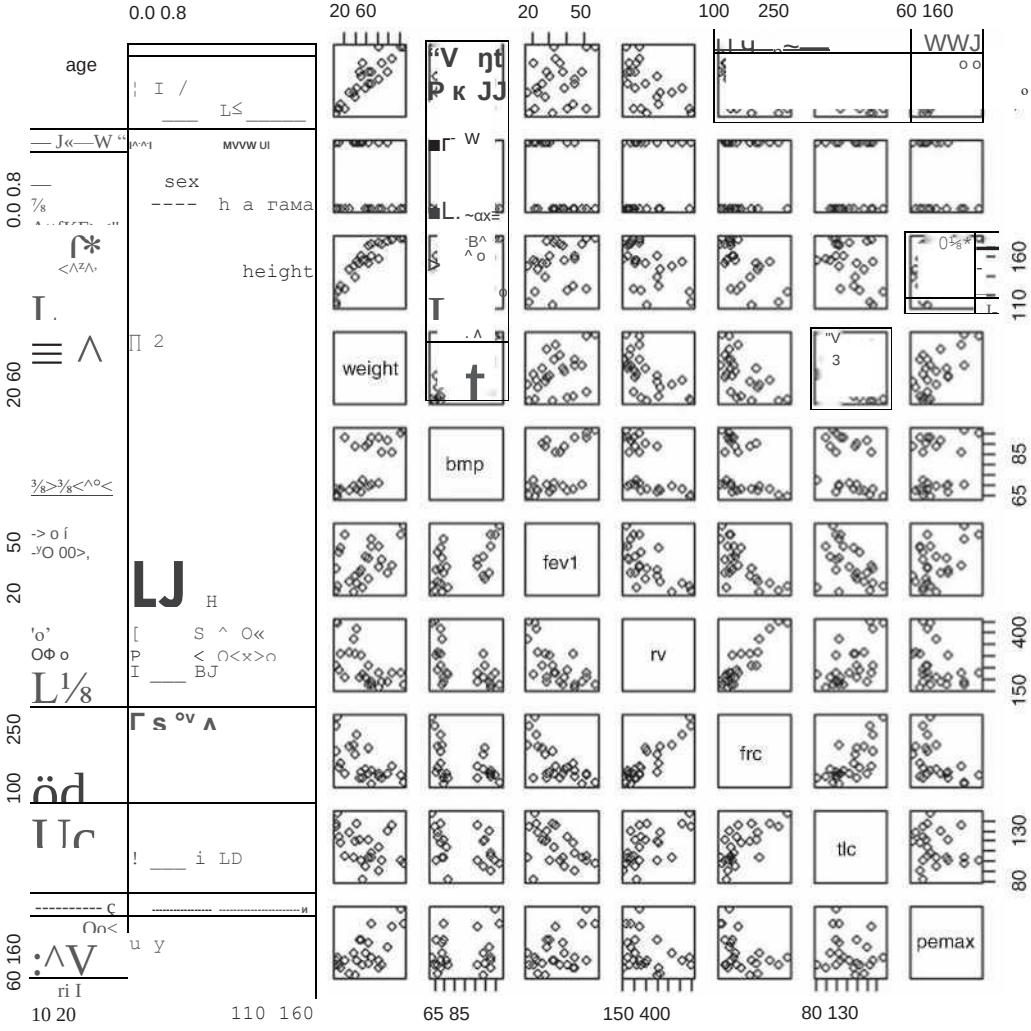


FIGURE 3.16 Plots for all 10 components of the cystic fibrosis data cystfibr.

Trasarea datelor multivariate: Pentru a obține diagrame de dispersie pe perechi între toate variabilele din cadrul de date, funcția R pairs() este utilizată în următorul segment de cod:

```
> par (mex=0,5)
> perechi (cystfibr, gap=0, cex.labels=0.9) # Ieșire Figura 3.17 .
```

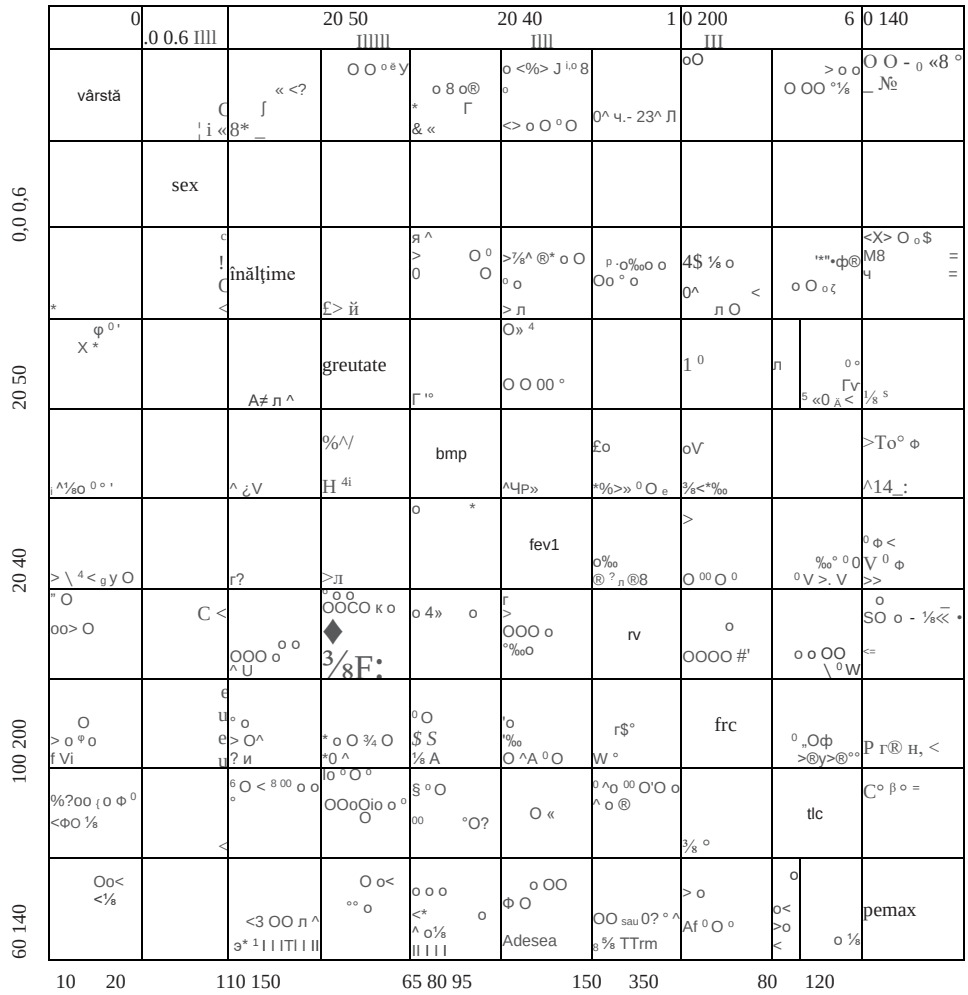


FIGURA 3.17 Grafice perechi îmbunătățite pentru datele de fibroză chistică cystfibr .

Aici, argumentele controlează aspectul intrigii:

- Parametrul grafică mex reduce distanța dintre linii în margini.
- gap (=0 , adică fără decalaj) elimină spațiul dintre subploturi.
- cex.labels (=0.9) scade dimensiunea fontului.

Observatii:

1. Deoarece parcelele individuale sunt mici, claritatea poate fi compromisă.
2. Cele 10 variabile din cadrul de date sunt vârsta , sexul , înălțimea , greutatea , bmp , fev1 , rv , frc , tlc și pemax .
   Acestea reprezintă următorii parametri:
   vârstă : vârstă, un vector numeric; vârsta în ani
   sex : sex, un cod vectorial numeric; 0: bărbat, 1: femeie
   inaltime : inaltime, un vector numeric; inaltime (cm)
   greutate : greutate, un vector numeric; greutate (kg)
   bmp : bmp, un vector numeric; procent de masă corporală (procent din normal)
   fev1 : fev1, un vector numeric; volumul expirator forțat

rv : rv, un vector numeric; volumul rezidual
frc : frc, un vector numeric; capacitatea reziduală funcţională
tlc : tlc, un vector numeric; capacitatea pulmonară totală
pemax : pemax, un vector numeric; presiunea expiratorie maxima

3. Cu toate acestea, sunt evidente unele caracteristici speciale și tendințe clare. De exemplu:
   (a) În cadrul de date cystfibr , există 10 variabile: vârstă , sex , înălțime , greutate , bmp , fev1 , rv , frc , tlc și pemax . Se pare că a zecea variabilă pemax poate fi considerată variabilă dependentă, celelalte nouă fiind considerate variabile independente.
   (b) În subplotul vârstă versus sex (Rândul 1, Coloana 2), deoarece există doar două sexe (M, F), toate datele sunt stivuite pe două grămezi: o coloană pentru M și o altă coloană pentru F.
   (c) În diagrama secundară greutate versus înălțime (Rândul 3, Coloana 4), există o tendință clară care arată că, printr-o pantă aproape liniară, greutatea este aproximativ direct proporțională cu înălțimea .
   (d) Acest cadru de date conține nume de variabile comune, cum ar fi vârsta , înălțimea , greutatea și așa mai departe. Pentru a elimina posibila confuzie, este prudent să vă asigurați că aceste obiecte cu nume identic nu sunt implicate în altă parte în mediul R din acest spațiu de lucru în aceeași sesiune de lucru.

Analiză ulterioară folosind modelarea biostatistică

Informații suplimentare pot fi obținute prin modelarea biostatistică a cadrului de date cystfibr . De exemplu, un model R liniar lm() poate fi utilizat pentru a corela datele. (În mod clar, pot fi utilizate alte modele biostatistice.)

Funcția lm() este utilizată pentru a se potrivi modelelor liniare. Poate fi utilizat pentru a efectua regresia, ANOVA cu un singur strat și analiza covarianței.

Forma pentru utilizarea lui lm() este

```
lm(formula, date, subset, ponderi, na.action,
    metoda = „qr", model = TRUE, x = FALS, y = FALS,
    qr = TRUE, singular.ok = TRUE, contraste = NULL, offset, ...)
```

Pentru un diagnostic clinic al fibrozei chistice folosind cadrul de date colectat, se poate lua în considerare parametrul **presiunii expiratorii maxime (MEP)** , pemax , care este o măsură a forței mușchilor respiratori și se obține prin expirarea cât mai puternică a pacientului. într-un mușiuc. *Valoarea maximă este aproape de capacitatea pulmonară totală* . Prin urmare, se poate defini ca funcția de corelare a modelului liniar să fie

```
> lm(pemax ~ vârstă + sex + înălțime + greutate + bmp + fev1 + rv + frc + tlc)
```

ceea ce înseamnă că pemax este descris folosind un model liniar care este aditiv în restul de nouă variabile. Ieșirea acestui segment de cod este

Apel:
lm(formula = pemax ~ vârstă + sex + înălțime + greutate + bmp +
fev1 + rv + frc + tlc)

Coeficienți:

(Interceptare)	vârstă	sex	înălțime	greutate
176,0582	-2,5420	-3,7368	-0,4463	2,9928
bmp	fev1	rv	frc	tlc
1,7449	1,0807	0,1970	-0,3084	0,1886

```
> # Folosind comanda rezumat, s-a > # obținut rezultate mai semnificative :
    > rezumat(lm(pemax ~ varsta + sex + inaltime + greutate + bmp +
    >                    fev1 + rv + frc + tlc))
    > # Ieşirea tabelului ANOVA :
```

Apel:

lm(formula = pemax ~ vârstă + sex + înălțime + greutate + bmp + fev1 + rv + frc + tlc)

Reziduuri:

Min	1 Î	Median	3Q	Max
-37.338	-11.532	1.081	.386	33.405

Coeficienți:

	Estima	Std.	valoarea t	Pr(>|t|)
(Intercepta)	176,0582	225,8912	0,779	0,448
vârstă	-2,5420	4,8017	-0,529	0,604
sex	-3,7368	15,4598	-0,242	0,812
înălțime	-0,4463	0,9034	-0,494	0,628
greutate	2,9928	2,0080	1.490	0,157
bmp	-1,7449	1,1552	-1.510	0,152
fev1	1,0807	1,0809	1.000	0,333
rv	0,1970	0,1962	1.004	0,331
frc	-0,3084	0,4924	-0,626	0,540
tlc	0,1886	0,4997	0,377	0,711

Eroare standard reziduală: 25,47 pe 15 grade de libertate R-pătrat multiplu: 0,6373, R-pătrat ajustat: 0,4197 F-statistică: 2,929 pe 9 și 15 DF, valoare p: 0,03195

Observatii:

1. Cu funcția lm() , nu a existat o cantitate mare de ieșire.
2. Cu ajutorul suplimentar al funcției summary() , s-au obținut mai multe rezultate.
3. Valorile *t* , în termeni absoluti, au variat de la 0,242 la 1,510; astfel, nu există o singură valoare *t* care să fie semnificativă din punct de vedere biostatistic. Cu toate acestea, în testul comun F, valoarea *p* a fost 0,03195, ceea ce a indicat că este un rezultat semnificativ. Motivul pentru acest rezultat este că testele *t* vor spune ceva despre ceea ce se întâmplă numai dacă o variabilă este eliminată în timp ce toate celelalte variabile sunt lăsate. Prin urmare, concluzia este că nu ar trebui inclusă o anumită variabilă. Nu este clar din punct de vedere biostatistic dacă vreo variabilă anume ar trebui eliminată pentru a forma un model redus.
4. R^2 neajustat de 0,6373 diferă semnificativ de R^2 ajustat de 0,4197. Acest lucru se datorează probabil numărului mare de variabile, 10, în raport cu numărul de DF pentru varianță, 15. Acest lucru este în concordanță cu faptul că prima este modificarea sumei reziduale a pătratelor în raport cu un model gol, în timp ce al doilea. este o modificare similară a variației reziduale.

R ajustat [16]se calculează după cum urmează:

R1 = Eroare standard reziduală = 25,47

Revenind la mediul R, pătratul R ajustat (notat ca R 2 în cod) poate fi găsit de următorul segment de cod:

```
>
>   R1 <- 25,47 # Aceasta este eroarea standard rezidual
> R2 <- 1 - R1^2/var(pemax)
>   R2 # Ieşirea R-pătratului ajustat
[1]  0,4197626
>
```

Întrebări de revizuire pentru Secțiunea 3.5

1. Definiți analizele de date univariate, bivariate și multivariate, dând un exemplu pentru fiecare.
2. (a) Cum sunt efectuate aceste analize în mediul R ?
   (b) Dați exemple de segmente de cod R pentru aceste analize.
3. (a) Ce se înțelege prin *analiza de regresie* ?
   (b) Cum este utilizată analiza de regresie în analiza datelor?
4. (a) Cum se efectuează analiza de regresie în mediul R ?
   (b) Furnizați exemple de funcții R utilizate pentru analiza de regresie.
5. (a) Rezumați cele două utilizări ale tabelului ANOVA în analiza datelor.
   (b) Pentru analiza datelor, sugerați un segment de cod R aplicabil.

Exerciții pentru Secțiunea 3.5

3. *Epidemiologia ritmului cardiac după tratamentul cu enalaprilat*

   Enalaprilatul este o modificare a medicamentului enalapril, un inhibitor al enzimei de conversie a angiotensinei (ACE) utilizat în tratamentul **hipertensiunii arteriale** și al unor tipuri de **insuficiență cardiacă cronică** . ACE crește tensiunea arterială prin constrângerea vaselor de sânge; prin urmare, **enalaprilatul** este numit *pro-medicament* **.**

   Setul de date heart.rate are 36 de rânduri și 3 coloane. Conține date pentru nouă pacienți cu insuficiență cardiacă congestivă înainte și la scurt timp după administrarea de enalaprilat, într-un aspect echilibrat în două sensuri. Acest cadru de date conține următoarele coloane:

16Folosind segmentul de cod R aici:
(a) Instalați pachetul ISwR .
(a) Apelați fișierele din acest pachet.
(b) Consultați setul de date „Presiunea arterială versus obezitate” bp.obese .
(c) Graficul obezității versus tensiunea arterială, diferențiind datele între bărbați și femei:

```
> install.packages ("ISwR")
> bibliotecă (ISwR)
> bp.obeze
> plot(bp~obese,pch = ifelse(sex = 1, "F", "M"), data = bp.obese)
```

hr un vector numeric; ritmul cardiac în bătăi pe minut
cu nivelurile de la 1 la 9
timp un factor cu niveluri 0 (înainte), 30, 60 și 120 (minute după administrare)

Folosind segmentul de cod R aici:

(a) Instalați pachetul ISwR .
(b) Apelați fișierul heart.rate din acest pachet.
(c) Inspectați setul de date.
(d) Pentru fiecare dintre cei nouă pacienți, graficați frecvența cardiacă medie în funcție de timp.

```
> install.packages(„ISwR")
> bibliotecă (ISwR)
> ritmul.inimii
> evalq(interaction.plot(time,subj,hr), heart.rate)
```

4. *Epidemiologie nutrițională: un exercițiu de analiză a datelor multivariate și analiza variației*

Mai întâi, iată câteva funcții R utile pentru această aplicație:

Funcția tapply() , din baza pachetului CRAN R , aplică o funcție fiecărei celule dintr-o matrice ragged; adică fiecărui grup nevid de valori date de o combinație unică a nivelurilor anumitor factori.

Funcția aov() , din statisticile pachetului CRAN R , se potrivește unui model ANOVA printr-un apel la lm pentru fiecare strat.

Investigația: creșterea în greutate la șobolanii de laborator

Iată o analiză interesantă într-o investigație în epidemiologia nutrițională. Datele provin dintr-un experiment pentru a studia creșterea în greutate a șobolanilor de laborator hrăniți cu patru diete diferite, distinse prin cantitatea de proteine (scăzută și ridicată) și după sursa de proteine (carne de vită și cereale).

Folosind segmentul de cod R aici:

(a) Instalați pachetul HSAUR .

```
> install.packages ("HSAUR")
```

(b) Apelați fișierul de creștere în greutate din acest pachet:

```
> date(„creștere în greutate", pachet = „HSAUR")
```

(c) Inspectați creșterea în greutate a setului de date:

```
> crestere in greutate
```

(d) Rezumați principalele caracteristici ale acestui set de date calculând mediile și abaterea standard, folosind funcția tapply():.

> tapply(creștere în greutate$creștere în greutate, listă(creștere în greutate$sursă,

+ creștere în greutate$tip), medie)

(e) Reprezentați grafic creșterea medie în greutate pentru cantitatea de factor proteic, arătând că câștigul pentru dieta bogată în proteine este mult mai mare decât pentru dieta săracă în proteine: > plot.design(weightgain)

(f) Rezumați variațiile celulare pentru a arăta că sunt relativ similare și că nu există o relație aparentă între media celulară și varianța celulară. Aceasta, la rândul său, arată că ipoteza de omogenitate a ANOVA pare a fi rezonabilă pentru acest set de date:
> tapply(creștere în greutate$creștere în greutate, listă(creștere în greutate$sursă,
+ creștere în greutate$tip), sd)

(g) Utilizați funcția R aov() pentru a efectua o ANOVA pentru setul de date; apoi examinați rezultatul analizei. În cele din urmă, utilizați funcția summary() pentru a produce un tabel ANOVA:
> wg_aov <- aov(creștere în greutate ~ sursă * tip, date = creștere în greutate) > wg_aov
> rezumat(wg_aov)

(h) Tabelul ANOVA rezultat arată că efectul principal al tipului este foarte semnificativ , confirmând ceea ce sa văzut în partea (e) a acestui exercițiu. Efectul principal al sursei nu este semnificativ, dar interpretarea ambelor efecte principale este complicată de interacțiunea sursei de tip X, care se apropie de semnificație la nivelul de 5%. Pentru a înțelege acest efect de interacțiune, reprezentați grafic creșterea medie în greutate pentru dietele sărace și bogate în proteine pentru fiecare nivel de sursă de proteine: carne de vită și cereale. Utilizați următorul segment de cod R:
> interaction.plot(weightgain$tip, weightgain$sursa,
+ creștere în greutate$creștere în greutate)

(i) Din graficul rezultat, se poate observa că pentru dietele sărace în proteine, utilizarea cerealelor ca sursă de proteine a dus la o creștere în greutate mai mare decât utilizarea cărnii de vită. Pentru dietele bogate în proteine, invers este adevărat, cu carne de vită/dieta bogată în proteine ducând la cea mai mare creștere în greutate. Obțineți estimările interceptării și efectele principale și ale interacțiunii prin extragerea acestora din potrivirea modelului folosind următorul segment de cod R :
> coef(wg_aov)

(j) Rețineți că modelul a fost prevăzut cu următoarele restricții:
$g_1 = 0$ (corespunzător cu Carne de vită) și $b_1 = 0$ (corespunzător cu High) deoarece contrastele de tratament au fost utilizate ca implicite. Acest lucru poate fi văzut din următorul segment de cod R :

$$\sum_i g_i = 0$$

```
> opțiuni ("contraste")
```

(k) Astfel, coeficientul pentru o sursă de 14,1 (în tabelul ANOVA) poate fi considerat o estimare a diferențelor $g_2 - g_1$. De asemenea, se poate folosi următoarea restricție:

$$\sum_i g_i = 0$$

prin următoarele calcule R ale coeficienților:

```
> coef(aov(creștere în greutate ~ sursă + tip + sursă:tip, date =
+              creștere în greutate, contraste = listă(sursă = contr.sum)))
```

5. ***Studiul Medicare privind poluarea aerului*** **(MCAPS;** Peng & Domonici, 2008 **)**

Pachetul MCAPS conține estimări de probabilitate maximă și variații biostatistice ale riscurilor log-relative specifice județului ale internărilor în spital pentru fiecare dintre bolile cardiovasculare și respiratorii asociate cu întârzierile 0, 1 și 2 ale expunerii la $PM_{2,5}$. ($PM_{2,5}$ și PM_{10} sunt măsuri de particule din **atmosferă** cu un diametru mai mic sau egal cu 2,5 și, respectiv, 10 micrometri nominal.) Pachetul conține, de asemenea, date despre poluarea aerului și vremea pentru cele șapte regiuni geografice utilizate.

Următoarele segmente de cod R încarcă pachetul, extrag seturile de date și examinează critic conținutul acestora.

(a) Rulați segmentul de cod.

(b) Explicați pe scurt rezultatul în fiecare etapă:

```
> install.packages ("MCAPS")
> bibliotecă (MCAPS)
> initMCAPS(„MCAPS")
> getData()
> estimări <-getData("estimates.subset")
> head(estimations[, c("CountyName", "rezultat", "beta", "var")])
> site-uri <- getData("siteList")
> cap (site-uri)
> apw <- getData("APWdata")
> șic <- apw[["17031"]]
> cap (chic)
```

6. *Fibroza chistică: analiza regresiei polinomiale*

Revenind la setul de date privind fibroza chistică din Exemplul 1 din Secțiunea 3.5, graficul pemax față de înălțime arată o neliniaritate considerabilă. Conform teoriei biostatisticii, pentru a testa această observație, poate fi instructiv adăugarea unui termen neliniar în înălțime, cum ar fi pătratul înălțimii: $(\text{înălțime})^2$.

(a) Pe această bază, furnizați un segment de cod R care adaugă efectul $(\text{înălțimii})^2$ la modelul de corelare.

(***Sugestie*** : Utilizați funcția R predict() pentru noul model; și newdata , care permite predicția valorilor pentru un set ales de predictori. Alegeți un set de înălțimi între 110 și 180 cm în pași de 2 cm.)

(b) Comentați succesul sau eșecul acestui nou model de corelație neliniară.

REFERINȚE

Adler, J. (2010). *R pe scurt: o referință rapidă pentru desktop.* Sebastopol, CA: O'Reilly Media, Inc.

Asociația Americană de Statistică. (nd). căutarea unui loc de muncă biostatistician. Preluat de pe http://jobs.amstat.org/jobs/4627784/biostatistician-1

ANOVA. (nd). Analiza varianței. Preluat de la http://en.wikipedia.org/wiki/ Analiza_variantei

Aragon, TJ (2011). *Epidemiologie aplicată folosind R* (*epir*). Berkeley, CA: UC Berkeley School of Public Health & San Francisco Department of Public Health.

Note IMC. (2012). *Indicele de masă corporală* . Recuperat de lahttp://en.wikipedia.org/wiki/Body_mass_index

Centrele pentru Controlul și Prevenirea Bolilor (CDC). (2005, ianuarie). Profilaxia post-expunere antiretrovială după consumul de droguri injectabile sau sexuale sau alte expuneri nonprofesionale la HIV în Statele Unite: Recomandări de la Departamentul de Sănătate și Servicii Umane din SUA. *Rapoartele MMWR Recommendations, 54* (RR-2), 1-20. Preluat de pe http://www.cdc .gov/mmwr/preview/mmwrhtml/rr5402a1.htm

CRAN. (nd). Rețeaua cuprinzătoare de arhive R. Preluat de pe http://cran.r-project.org/ Dalgaard, P. (2002). *Statistici introductive cu R* (Springer Statistics & Computing Series). New York, NY: Springer Science+Business Media.

Daniel, WW (2005). *Biostatistică: o bază pentru analiză în științele sănătății.* New York, NY: John Wiley & Sons.

Everitt, BS și Hothorn, T. (2006). *Un manual de analiză statistică folosind R.* Boca Raton, FL: Chapman & Hall/CRC.

Mittal, HV (2011). *R carte de bucate grafice* . Birmingham, Marea Britanie: PACKT Publishing Ltd.

Murrell, P. (2006). *grafica R.* Seria Informatică și Analiză a Datelor. Boca Raton, FL: Chapman & Hall/CRC.

Peng, RD, & Domonici, F. (2008). *Metode statistice pentru epidemiologia mediului cu R: Un studiu de caz în poluarea aerului și sănătatea.* Springer Folosește R! Serie. New York, NY: Springer Science+Business Media.

Seefeld, K. și Linder, L. (2005). *Statistici folosind R cu exemple biologice.* Durham, NH: Departamentul de Matematică și Statistică, Universitatea din New Hampshire. (De asemenea, afiliat cu Departamentul de Nefrologie și Centrul de Cercetare în Biostatistică, Tufts-NEMC, Boston, MA.) Preluat de la http://cran.r-project.org/doc/contrib/Seefeld_StatsRBio.pdf

Statistics Canada. (23 iulie 2013). Prelucrarea datelor. Preluat de la http://www.statcan.gc.ca/edu/ power-pouvoir/ch3/5214785-eng.htm#a1

Teetor, P. (2011). *Carte de bucate R: rețete dovedite pentru analiza datelor, statistici și grafică.* Sebastopol, CA: O'Reilly Media.

Venables, WN, Smith, DM și echipa principală de dezvoltare R. (2004). *O introducere în R.* Bristol, Marea Britanie: Network Theory, Ltd.

Verzani, J. (2005). *Folosind R pentru statistici introductive.* Boca Raton, FL: Chapman & Hall/CRC.

Virasakdi, C. (2011). *Analiza datelor epidemiologice folosind R și Epicalc.* Songkhla, Thailanda: Unitatea de Epidemiologie, Prince of Songkla University.

ANEXĂ: DOCUMENTAȚIE PENTRU FUNȚIA intrigii

plot {grafica} R Documentatie

Graficul X–Y generic DESCRIERE

Funcție generică pentru reprezentarea grafică a obiectelor R. Pentru mai multe detalii despre argumentele parametrilor grafici, vezi alin .

Pentru diagrame de dispersie simple, va fi utilizat plot.default . Cu toate acestea, există metode de grafică pentru multe obiecte R , inclusiv funcții , data.frames , obiecte de densitate și așa mai departe. Folosiți metode (complot) și documentația pentru acestea.

UTILIZARE

grafic (x, y, ...)

ARGUMENTE

x Coordonatele punctelor din grafic. Alternativ, poate fi furnizată o singură structură de reprezentare, funcție sau orice obiect R cu o metodă de reprezentare .

y Coordonatele y ale punctelor din grafic; opțional, dacă x este o structură adecvată.

... Argumente care trebuie transmise metodelor, cum ar fi parametrii grafici (vezi alin.). Multe metode vor accepta următoarele argumente: type Ce tip de parcelă trebuie desenat. Tipurile posibile sunt

- „p" pentru puncte
- „l" pentru linii
- „b" pentru ambele
- „c" pentru liniile numai partea lui „b"
- „o" pentru ambele „supraplot"
- „h" pentru „histogramă" ca (sau „de înaltă densitate") linii verticale
- „s" pentru treptele scărilor
- „S" pentru alți pași (vezi „Detalii" de mai jos)
- „n" pentru nici un complot

Toate celelalte tipuri dau un avertisment sau o eroare; folosind, de exemplu, type = "punkte" fiind echivalent cu tip = "p" pentru compatibilitatea S. Rețineți că unele metode (de exemplu, plot.factor), nu acceptă acest lucru. principal

Un titlu general pentru intriga; vezi titlul. sub

Un subtitrare pentru intriga; vezi titlul. xlab

Un titlu pentru axa x; vezi titlul.

ylab

Un titlu pentru axa y; vezi titlul.

asp

Raportul de aspect y/x; vezi plot.window.

DETALII

Cele două tipuri de pași diferă în preferința lor xy. Trecând de la *(x1,y1)* la *(x2,y2)* cu *x1 < x2* , type = "s" se mișcă mai întâi pe orizontală și apoi pe verticală, în timp ce tip = "S" se mișcă invers.

Vezi, de asemenea , **plot.defvault, plot.formula** și alte metode; **puncte** , **linii** , **alin**

.

Pentru trasarea X–Y–Z, consultați **contur** , **persp** și **imagine** .

Exemple

```
cere(statistici)
parcela (masini)
linii (mai joase (mașini))
plot(sin, -pi, 2*pi) # vezi ?plot.function
## Grafic de distribuție discretă:
plot(table(rpois(100,5)), tip = „h", col = „roșu", lwd=10,
      main="rpois(100,lambda=5)")
## Cuantile simple/ECDF; vezi ecdf() {bibliotecă(stats)} pentru una mai bună:
plot(x <- sort(rnorm(47)), type = "s", main = "plot(x, type = \"s\")" ) puncte(x, cex
= .5, col = „roșu închis")
```

PATRU

Grafică folosind R

INTRODUCERE

Până în acest moment, ați fost introdus în multe ieșiri grafice în mediul R , toate pe o bază ocazională, ad-hoc. În mod clar, grafica este o caracteristică importantă și versatilă în biostatistica epidemiologiei și a sănătății publice. Acest capitol investighează astfel de facilități grafice în detaliu.

R poate fi descrisă în termeni de două sisteme (Dalgaard , 2002 ; Mittal, 2011; Murrell, 2006; Echipa centrală de dezvoltare Venables, Smith și R , 2005) : grafică de bază (sau tradițională) și grafică grilă.

Sistemul grafic R poate fi considerat ca fiind format din patru niveluri:

1. **Pachete grafice** : listate în cea mai mare parte pe site-ul CRAN
2. **Sisteme grafice** : inclusiv *grafica* și *grila* , care vor fi discutate în detaliu în acest capitol
3. **Motoare grafice:** cunoscute ca *grDevices* ; acestea permit utilizatorilor să se ocupe de aspecte precum tipurile și culorile fonturilor, formatele de ieșire și altele asemenea
4. **Pachete de dispozitive grafice:** inclusiv pachete grafice *suplimentare* , care oferă detalii despre ieșirile grafice

Sistemul de bază, împreună cu pachetele grafice construite pe el, oferă majoritatea funcțiilor de nivel înalt. Excepție este pachetul de zăbrele, care oferă parcele complete bazate pe sistemul de grilă. Atât sistemele de bază, cât și cele de rețea pot fi utilizate în moduri batch sau interactiv; cu toate acestea, acesta din urmă este mai productiv. Pentru utilizare interactivă, la momentul pornirii, R inițiază un *driver de dispozitiv grafic* care deschide o fereastră grafică pentru afișarea graficelor interactive. În Microsoft Windows, comanda este pur și simplu windows() .

De îndată ce driverul de dispozitiv este pornit, comenzile de trasare R pot fi folosite pentru a produce grafice și pentru a crea diferite afișaje. Comenzile de trasare disponibile sunt clasificate în trei grupuri:

Grupa I: Funcții de nivel înalt care creează o diagramă nouă, completă pe dispozitivul grafic (cu axe, linii, puncte, etichete, titluri etc.)

Grupa II: funcții de nivel scăzut care adaugă informații la o diagramă existentă (cum ar fi linii, puncte suplimentare, etichete etc.)

Grupa III: Funcții interactive care adaugă în mod interactiv informații la, sau elimină informații dintr-o parcelă existentă folosind un dispozitiv de indicare, cum ar fi un touchpad sau un mouse

R de mai sus sunt cunoscute ca *grafică de bază* . Un subsistem, care coexistă cu baza, este un pachet numit grid graphics ; acesta conține zăbrele de pachet , care oferă coduri pentru a produce diagrame pe mai multe niveluri similare cu sistemul Trellis din S.

Alegerea sistemului

Deoarece R poate produce diagrame complete cu un singur apel de funcție, alegerea sistemelor grafice de utilizat depinde doar de tipul de diagramă necesar. Dacă este necesar să adăugați ieșiri suplimentare la o diagramă inițială, este important să cunoașteți sistemul specific care a fost utilizat pentru a produce graficul original: în general, aceleași sisteme grafice ar trebui să fie utilizate pentru a adăuga ieșiri la diagramele existente (deși există excepții). Sistemul de grilă oferă de obicei mai multă flexibilitate.

Pachete

Funcțiile din R sunt organizate în termeni de *pachete* , ceea ce permite:

- Încărcarea numai a pachetelor care conțin funcțiile necesare; aceasta necesită mai puțină memorie și, astfel, permite lui R să ruleze mai rapid.
- Utilizarea funcțiilor pe care alții le-au scris și încărcate în pachet; acum există sute de aceste pachete contribuite.

Pentru a descărca și instala pachete dintr-o sesiune R , utilizați:

```
> install.packages(„NUMELE PACHET")
```

și

```
> update.packages(„NUMELE PACHET")
```

La instalare, pentru a încărca funcțiile din acel pachet, utilizați:

```
> bibliotecă(„NUMELE PACHET")
```

Note:

1. Sistemul grafic de bază este furnizat de pachetul grafic , care este instalat și încărcat, implicit, într-o instalare standard a R .
2. Sistemul Grid Graphics este furnizat de pachetul grid , care, împreună cu pachetul Lattice , este de asemenea instalat implicit atunci când este încărcat sistemul Grid Graphics.
3. Pentru a automatiza încărcarea pachetelor, utilizați:
   > ajutor (pornire)
4. Pentru a obține ajutor pentru orice funcție FUNCTION , utilizați:
   ```
   > ajutor (FUNCTIE)
   ```
   Pentru mai mult ajutor, încercați una sau toate dintre următoarele:
   ```
   > ajutor(ajutor) # Pentru ajutor suplimentar
   > example() # Pentru unele exemple pentru a rula funcția studiată
   > help.start() # Vă duce la ajutorul suplimentar disponibil
   > help.search() # Pentru a localiza o funcție cu un scop special
   ```

> vignette() *# Vignetele au ajutor mai extins*

În cele din urmă, pagina de pornire R are un ghid pentru forumul R-help , căruia îi trimiteți întrebări pentru asistență online.

4.1 GRAFICA DE BAZĂ (SAU TRADIȚIONALĂ).

Funcții de nivel înalt

lui R generează un grafic complet al datelor care au fost transmise ca argumente ale funcției. Dacă nu se solicită altfel, axele, etichetele și titlurile sunt generate. O comandă de nivel înalt pornește un nou grafic, ștergând graficul curent dacă este necesar.

FUNȚIA plot()

Aceasta este cea mai folosită comandă de trasare; este o funcție *generică* . Adică, tipul de diagramă produs depinde de *clasa* sau tipul primului argument al funcției. Următoarele sunt câteva exemple tipice.

> **Exemplul 4.1**

> complot(x)

Dacă *x* este o serie de timp, plot(x) va produce o diagramă de serie de timp. Dacă *x* este un vector numeric, această funcție va produce un grafic al valorilor din vector față de indicii lor respectivi din vector. Dacă *x* este un vector de numere imaginare, această funcție va produce o diagramă a părților imaginare versus reale ale componentelor vectoriale.

Ca o ilustrare a funcției în mediul R , luați în considerare:

```
> x <- 1:25 # Fie x vectorul primelor 25 de numere naturale
> x # Ieșirea x :
[1]  1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
> plot(x) # Ieșire Figura 4.1 .
```

TRACTARE MULTIVARIAT DE DATE

Pentru trasarea datelor multivariate în R , există mai multe funcții speciale: pairs() , coplot() , stars() și mosaicplot() . Unele sunt descrise aici, cu exemple.

pairs(X) , unde X este o matrice numerică sau un cadru de date, emite o diagramă de dispersie pe perechi a variabilelor definite de coloanele lui X. În acest grafic, fiecare coloană a lui X este reprezentată pe fiecare altă coloană a lui X, iar n (n – 1) diagramele, unde n este numărul de variabile, sunt aranjate într-o matrice cu scalele grafice constante pe rândurile și coloanele sale.



FIGURA 4.1 > plot(x).

> **Exemplul 4.2:** Un exemplu numeric simplu al perechilor de funcții (X)

```
> x <- c(1, 2, 3, 4, 5)
> y <- c(1, 4, 9, 16, 25) # Notă: yi = xi2, i = 1,2,3,4,5
> X <- data.frame(x, y)
> X
```

	x	y
1	1	1
2	2	4
3	3	9
4	4	16
5	5	25

```
> perechi (X)
> # Ieșiri: Figura 4.2.
```

> **Exemplul 4.3:** Ieșirea unui coplot() simplu

```
> ## Cutremurele din tranșeul Tonga
> coplot(lat ~ long | depth, data = quakes) # Ieșire: Figura 4.3.
```

Dacă *c* este un factor, *a* este reprezentat grafic în raport cu *b* pentru valorile lui *c* în interval. Numărul și poziția intervalelor pot fi controlate cu argumentul dat.values= în coplot() . De asemenea, se pot folosi două variabile date cu comanda coplot(a ~ b | c + d) , care generează diagrame de dispersie ale *a* versus *b* pentru fiecare interval de condiționare comun al *c* și *d.*

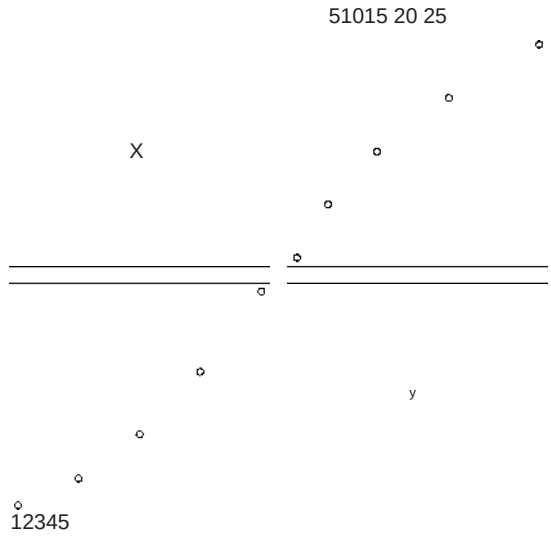


FIGURA 4.2
perechi(x).

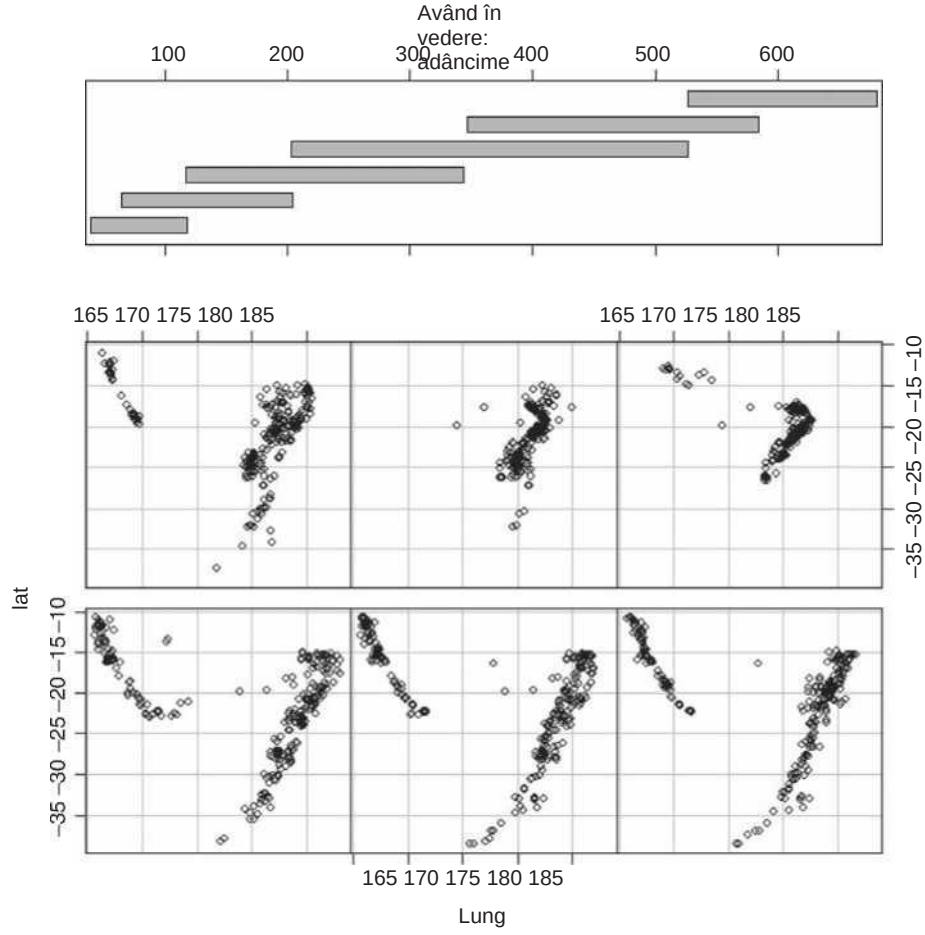


FIGURE 4.3 > coplot(a ~ b | c).

> **Exemplul 4.4:** pentru persp(x, y, z, ...)

Acest exemplu este extras din pagina online „Documentare" a funcției persp() . [1]

```
> require(grDevices) # pentru trans3d
> ## Exemple în demo(persp)
>
> # (1) Suprafața Matematică Obligatorie.
> # Funcția sinc rotită.
>
> x <- seq(-10, 10, lungime= 30)
> y <- x
> f <- function(x,y) { r <- sqrt(x^2+y^2); 10 * sin(r)/r}
> z <- exterior(x, y, f)
> z[is.na(z)] <- 1
> op <- par(bg = "alb")
> persp(x, y, z, theta = 30, phi = 30, expand = 0,5,
> col = „albastru deschis")
> # Ieșire: Figura 4.4.
>
> persp(x, y, z, theta = 30, phi = 30, expand = 0,5,
+ col = "albastru deschis",
+ ltheta = 120, nuanță = 0,75, tip tick = „detaliat",
+ xlab = "X", ylab = "Y", zlab = "Sinc(r)")
> # Ieșire: Figura 4.5.
```

Figura 4.5 > persp().

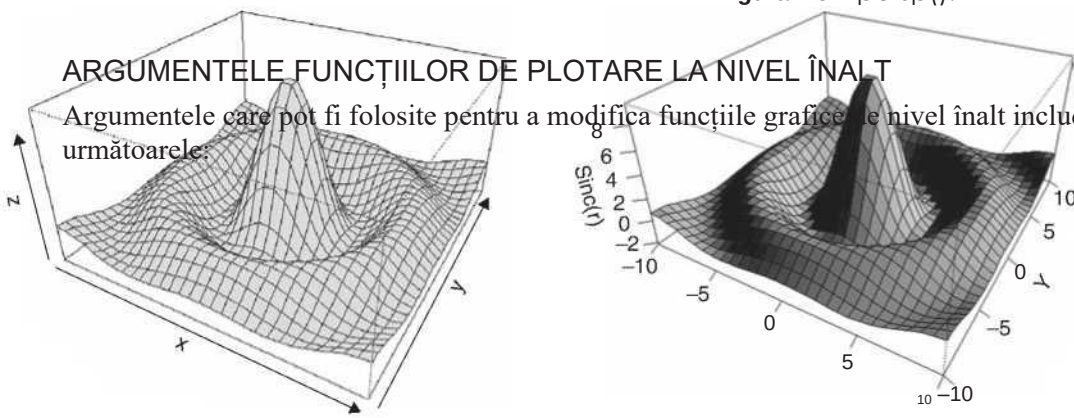


FIGURE 4.4 > persp().

ARGUMENTELE FUNCȚIILOR DE PLOTARE LA NIVEL ÎNALT

Argumentele care pot fi folosite pentru a modifica funcțiile grafice de nivel înalt includ următoarele:

[1] CRAN site for image(){graphics}:
http://127.0.0.1:27481/library/graphics/html/image.html

add=TRUE Acest argument suplimentar va face ca funcția să acționeze ca o funcție de nivel scăzut, suprapunând graficul pe diagrama curentă.
add=FALSE Acest argument suprimă generarea de axe și poate fi utilizat atunci când adăugați axe personalizate cu funcția axis() .
log ="x", log="y", log"xy" Aceste argumente fac ca axele respective să fie logaritmice.
tip= Acest argument controlează tipurile de diagrame produse, după cum urmează:

"p"	trasează	puncte individuale
"l"	trasează	liniile
"b"	trasează	punctele legate prin linii
"o"	suprapuse	cu linii
"h"	trasează linii verticale de la puncte la axa zero	
"s" , "S"	trasează funcțiile pasului, cu partea de sus/de jos, respectiv, a verticalei definirea punctului	
"n"	nici un complot	

xlab/ylab=(string)	Inserează etichete de axe pentru axele *x* și respectiv *y* .
main/sub=(string)	Inserează titlul figurii/subtitlul, respectiv, plasând textul în partea de sus a parcelei/chiar sub axa *x* .

Funcții de plotare la nivel scăzut

Graficul la nivel scăzut poate fi utilizat pentru a adăuga informații suplimentare, cum ar fi puncte, linii, text și așa mai departe, la un afișaj existent. Câteva funcții utile de trasare la nivel scăzut sunt:

puncte (x, y) Desenează o succesiune de puncte la coordonatele specificate. Caracterele specificate sunt trasate, centrate la coordonate.
linii (x, y) Adaugă puncte sau linii de legătură la graficul existent.

Un tip=argument poate fi folosit cu aceste funcții, fiind implicit „ p ” pentru puncte() și „ l ” pentru linii().

text(x, y, labels, …) Adaugă text la o diagramă în punctele x, y. labels este de obicei un vector întreg sau caracter; labels[i] este reprezentat grafic în punctele (x[i], y[i]) . Valoarea implicită este 1:length(x) . Astfel, pentru a trasa un set de puncte etichetate, se poate folosi următoarea secvență:
> plot(x, y, type="n"; text(x, y, names)

Aici, type="n" suprimă punctele, dar setează axele, în timp ce funcția text() furnizează caractere speciale, specificate de numele vectorului de caractere pentru puncte.

abline() adaugă una sau mai multe linii drepte prin graficul curent.
abline(a, b) adaugă o linie dreaptă cu interceptarea *a* și panta *b* la graficul existent.
abline(v=x) specifică coordonatele *x* pentru lățimile liniilor verticale care trebuie să parcurgă un grafic. abline(h=y) specifică coordonatele *y* pentru înălțimile liniilor orizontale care urmează să treacă pe o diagramă.
abline(lm.obj) specifică o listă, lm/obj , cu o componentă coeficientă de lungime 2, care sunt luate ca intersecție și, respectiv, panta.
axis(side, …) adaugă o axă la graficul existent pe partea specificată de primul argument (1–4, numărând în sensul acelor de ceasornic de jos).

Alte argumente controlează pozițiile axelor în interiorul sau lângă diagramă și bifează pozițiile și etichetele. Este util pentru adăugarea de axe după prima apelare a plot() cu argumentul axes=FALSE .

title(principal, sub) adaugă un titlu principal în partea de sus a diagramei curente într-un font mare și, opțional, un subtitrare în partea de jos (într-un font mai mic).

legend(x, y, legend, …) adaugă o legendă la graficul existent ca etichete de poziție în legenda vectorului de caractere . De asemenea, trebuie specificat cel puțin un argument suplimentar *v* (un vector cu aceeași lungime ca legenda), cu valorile corespunzătoare ale unității de trasare, după cum urmează:

legend(, fill=v)	specifică culorile pentru casetele umplute.
legend(, col=v)	specifică culorile în care vor fi desenate linii sau puncte.
legend(, lty=v)	specifică stilurile de linii.
legend(, lwd=v)	specifică lățimea liniilor.
legend(, pch=v)	specifică caracteristicile de trasare (un vector de caractere).

> **Exemplul 4.5:** legendă(x, y, legendă, ...)

Această funcție poate fi folosită pentru a adăuga legende în parcele. Forma sa documentară este:

```
legenda(x, y = NULL, legenda, umplere = NULL, col = par("col"), border="black",
        lty, lwd, pch,
        unghi = 45, densitate = NULL, bty = "o", bg = par("bg"), box.lwd =
        par("lwd"), box.lty = par ("lty"), box.col = par ("fg"),
        pt.bg = NA, cex = 1, pt.cex = cex, pt.lwd = lwd, xjust = 0, yjust = 1,
        x.intersp = 1, y.intersp = 1, adj = c(0, 0,5) , text.width = NULL, text.col =
        par("col"), merge = do.lines && has.pch, trace = FALSE, plot = TRUE, ncol
        = 1, horiz = FALSE, title = NULL, inset = 0, xpd, title.col = text.col, title.adj =
        0,5, seg.len = 2)
```

Pentru a ilustra utilizarea sa, pentru o anumită parcelă existentă, legende vor fi adăugate la fiecare dintre cele nouă poziții majore de pe parcelă:

```
>  x <- 0:1
>  y <- sin(x)
>  plot(x, y, type='n')
> # Ieșire: Figura 4.6.
> legenda("dreapta jos", "(x,y)", pch=1, titlu="dreapta jos")
> legendă(„de jos", „(x,y)", pch=1, titlu="de jos")
> legenda("stânga jos", "(x,y)", pch=1, titlu="stânga jos")
> legenda("stânga", "(x,y)", pch=1, titlu="stânga")
> legendă("topleft","(x,y)", pch = 1, title = "topleft, inset = 0,05",
+ insert = .05)
> legendă ("sus", "(x,y)", pch=1, title="top")
> legendă ("topright", "(x,y)", pch=1, title="topright, inset = 0,02",
+ inserție = 0,02)
```

```
> legendă(„dreapta", „(x,y)", pch=1, titlu="dreapta")
> legendă(„centru", „(x,y)", pch=1, titlu="centru")
>   # ieșire: Figura 4.7.
```

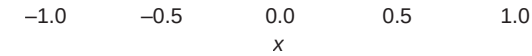


FIGURE 4.6 > plot().

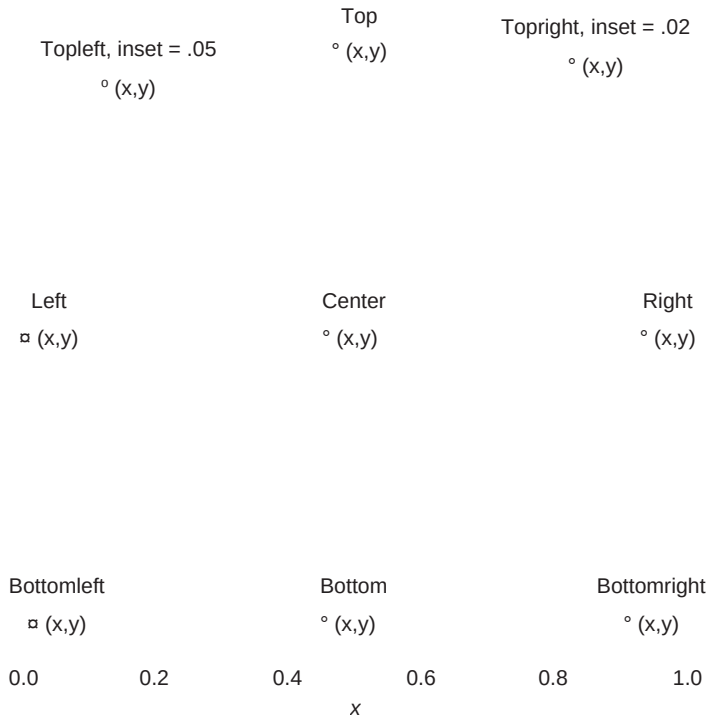


FIGURE 4.7 > legend(x, y, legend, ...).

ANOTAȚIE MATEMATICĂ

În mediul R , pentru a scoate o expresie matematică dată ca grafic, se poate folosi funcția plotmath() , din pachetul grDevices . Caracteristicile acestei funcții pot fi găsite în demo(plotmath) sau help(plotmath) , printre altele. Urmează câteva exemple.

> **Exemplul 4.6:** Afișarea expresiilor matematice[17]

O expresie matematică trebuie să respecte regulile normale de sintaxă pentru orice expresie R , dar este interpretată după reguli foarte diferite decât pentru expresiile R normale . Cu toate acestea, este posibil să se producă multe simboluri matematice diferite, să genereze superscripte sau indice, să producă fracții și așa mai departe.

Observați modurile de exprimare R pentru o expresie matematică tipică care include semnul de însumare, așa cum este dat de următorul segment de cod R :

```
> complot (1:10, 1:10)
> text(4, 7, expresie(bar(x) == suma(frac(x[i], n), i==1, n)))
> text(4, 6.4, „expresie(bar(x) == suma(frac(x[i],n), i==1, n))",
+ cex = .8) # Ieșire: Figura 4.8 .
```

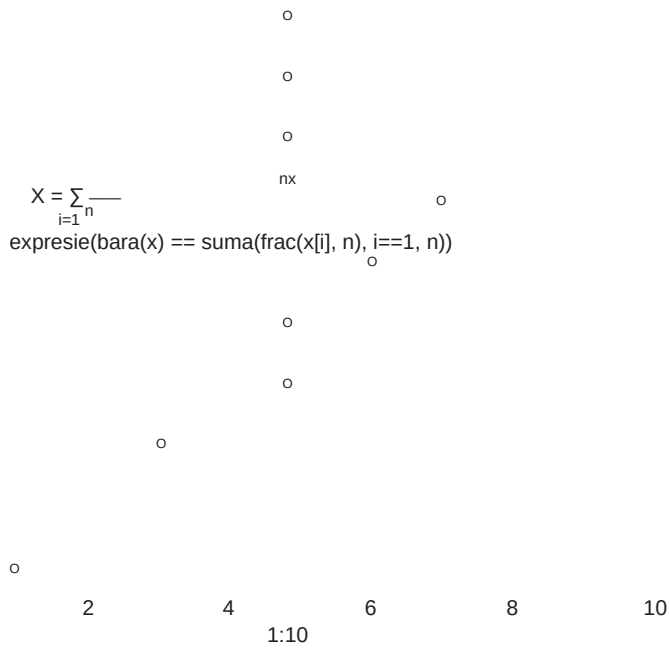


FIGURA 4.8 Afișarea unei expresii matematice.

Interacțiunea cu grafica

R care permit utilizatorilor să adauge sau să elimine informații dintr-o parcelă (folosind un mouse) includ locator(n = 512, tip = "n", ...):

[17] Site CRAN pentru plotmath{grDevices}: http://127.0.0.1:10508/library/grDevices/html/plotmath.html

locator(n = 512, tip = "n", ...). Această funcție așteaptă ca utilizatorul să selecteze anumite locații pe parcela existentă folosind butonul stâng al mouse-ului. Acest proces continuă până când au fost selectate *n puncte sau este apăsat un alt buton al mouse-ului.* Argumentul tip permite trasarea în punctele selectate și are același efect pentru grafica de nivel înalt. Valoarea implicită nu este reprezentarea grafică. Această funcție returnează locațiile punctelor selectate. Când este apelată fără un argument, această funcție are o valoare implicită mare (512).

identific(x, y, etichete) Această funcție permite evidențierea oricăror puncte prin x și y , folosind butonul stâng al mouse-ului prin trasarea componentei corespunzătoare a etichetelor din apropiere. Returnează indicii punctelor selectate.

Utilizarea parametrilor grafici

Se pot schimba multe aspecte ale unui afișaj grafic folosind parametrii grafici care controlează caracteristici precum tipurile și lățimile liniilor, culorile, aranjamentele figurilor, manipulările de text și așa mai departe. Fiecare parametru grafic are un nume; de exemplu, col controlează culorile și o valoare (de exemplu, un număr de culoare).

Parametrii grafici pot fi setați fie permanent (afectând toate funcțiile grafice care accesează dispozitivul existent), fie temporar (afectând doar un singur apel grafic).

MODIFICĂRI PERMANENTE: FUNCȚIA par()

par() , în pachetul grafic , poate fi folosit pentru a interoga sau pentru a seta parametrii grafici. Parametrii pot fi setati prin specificarea lor ca argumente la par în forma tag = valoare sau prin trecerea lor ca o listă de valori etichetate. Câteva exemple tipice sunt:

par() Fără argumente, această funcție returnează o listă cu toți parametrii grafici și valorile acestora pentru dispozitivul grafic existent.

par(c("lty", "lwd", "col")) Cu un argument vector caracter, această funcție returnează doar parametrul grafic numit, care în acest caz sunt tipul liniei, lățimea liniei și, respectiv, culoarea liniei.

par(col=3, lty=1) Cu argumentele numite setarea valorilor parametrilor grafici numiți, această funcție returnează doar valorile originale ale parametrului sub formă de listă.

Prin salvarea rezultatului par() la efectuarea modificărilor, se pot restabili valorile inițiale când trasarea este completă.

> **Exemplul 4.7:** Funcția par()

```
> par1 <- par(mfrow = c(3, 3), # 3 × 3 imagini pe o singură parcelă
>             col=3, pty = "s")   # cu culoare, regiune de reprezentare pătrată,
>                                 # independent de dimensiunea dispozitivului

> # La sfârșitul trasării, resetați la setările anterioare:
> p a r (p a rl)
```

MODIFICĂRI TEMPORARE PRIN ADĂUGAREA DE ARGUMENTE LA FUNCȚII GRAFICE

Adăugând parametri grafici la o funcție grafică, cum ar fi plot() , rezultatul are același efect ca și transmiterea argumentelor pentru a efectua modificări exprimate într-o funcție par() , cu excepția faptului că modificările indicate vor dura doar pe durata graficului. apel de funcție. Astfel, de exemplu,

```
> plot(x, y, pch=20)
```

va produce o diagramă de împrăștiere folosind „gloanțe" (cercuri solide mici) ca caracter de trasare , fără a face modificări în caracterul implicit pentru orice diagrame viitoare.

Lista parametrilor pentru grafică

Mulți parametri grafici utili sunt subliniați în următoarele subsecțiuni.

ELEMENTE GRAFICE

O listă completă de parametri pentru grafică poate fi găsită cu ușurință în următoarele două site-uri de documentație a funcției R , care pot fi apelate după cum urmează:

```
> ?par
> ?puncte sau ?pch
```

1. par() {grafică} poate fi folosit pentru a seta sau a interoga parametrii grafici. Parametrii pot fi setati prin specificarea lor ca argumente pentru par() .
2. points() {grafică} este o funcție generică pentru desenarea unei secvențe de puncte la coordonatele specificate. Caracterele specificate sunt trasate și centrate la coordonate. Documentația o include pe cea a pch() .

BIFĂRI ȘI AXE

Pentru diagramele de nivel înalt care necesită axe, se pot construi axe cu funcția grafică de nivel scăzut axis() . Cele trei componente ale unui set de axe sunt:

1. **Linia axei** , cu stilul liniei controlat de parametrul grafic lty (linie ty pe)
2. Bifurile care marchează diviziunile unităților de-a lungul liniei **axei**
3. Etichetele **de bifare** care marchează unitățile

Exemple de aceste caracteristici sunt:

```
xaxs="r", yaxs="i"
```

Cel de mai sus este stilul de calcul al intervalului de axă care trebuie utilizat pentru axele *x* și *y* . Valorile posibile sunt „r" și „i" . Stilurile sunt în general controlate de intervalul de date sau xlim , dacă este dat. Stilul „r" (regulat) extinde mai întâi intervalul de date cu 4% la fiecare

sfârșit și apoi găsește o axă cu etichete frumoase sau atractive din punct de vedere vizual care se încadrează în intervalul extins. Stilul „i" (intern) găsește doar o axă cu etichete frumoase care se încadrează în intervalul de date original.

tck=0,5

Cele de mai sus reprezintă lungimea semnelor de bifă ca o fracțiune din cea mai mică dintre lățimea sau înălțimea regiunii de trasare. Dacă tck ≥ 0,5 , este interpretat ca o fracțiune a laturii relevante, deci dacă tck = 1, sunt trasate linii de grilă. Setarea implicită este tck = NA .

lab=c(3, 6, 12)

Un vector numeric de forma c(x, y, len) care modifică modul implicit de adnotare a axelor. Valorile lui x și y dau numărul (aproximativ) de marcaje pe axele *x* și *y* , iar len specifică lungimea etichetei. Valoarea implicită este c(5, 5, 7) .

MARJILE FIGURA

În R , o singură parcelă, cunoscută sub numele de *figură* , cuprinde o **regiune a graficului** înconjurată de margini care pot conține titluri, subtitrări, etichete de axe și așa mai departe. Această regiune este de obicei delimitată de axe. Parametrii care controlează aspectul figurii includ:

mai=c(1.5, 1.0, 1.5, 1.0) Acești patru parametri sunt lățimea la marginile de jos, stânga, sus și, respectiv, din dreapta, măsurate în inci (Figura 4.9).
mar=c(4, 3, 4, 3) Similar cu mai , cu unitatea de măsură în linii de text.

Aceste două seturi de parametri sunt echivalente deoarece specificarea unui set schimbă valoarea celuilalt. Pentru mai multe detalii sau alte seturi de parametri pentru marginile figurilor, consultați documentația paragraful {grafice} din CRAN.

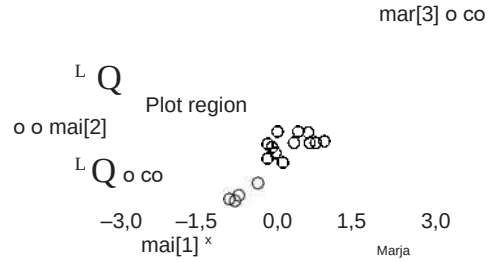


FIGURA 4.9 mar = c(a, b, c, d) : Un vector numeric de forma c(jos, stânga, sus, dreapta) , care dă dimensiunea marginii specificată în inci.

IEȘIRI MULTIPLE DE FIGURE

În R , se poate crea o matrice *n* × *m* de figuri într-o pagină, fiecare figură având propriile sale margini și cu matricea de figuri înconjurată opțional de o margine exterioară sau de margine. În documentația para {grafică} , această configurație poate fi specificată după cum urmează:

mfcol, mfrow Un vector de forma c(nr, nc) . Cifrele ulterioare vor fi desenate într-o matrice nr -by- nc pe dispozitiv pe *coloane* (mfcol), sau , respectiv, *rânduri* (mfrow).

oma, omi Un vector de forma c(jos, stânga, sus, dreapta) care indică dimensiunea marginilor exterioare în linii de text, sau respectiv inci.

O matrice cu 3 rânduri × 2 coloane este prezentată în Figura 4.10 .

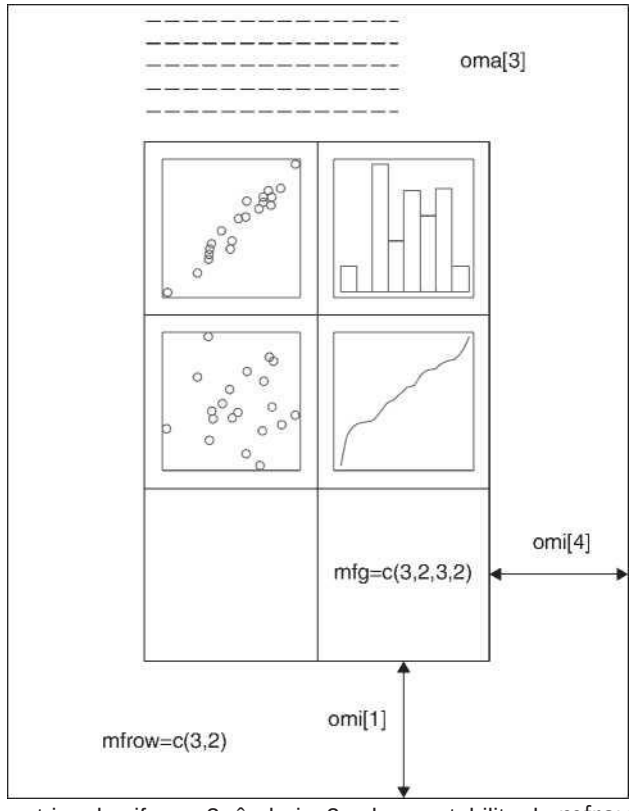


FIGURA 4.10 O matrice de cifre cu 3 rânduri x 2 coloane stabilite de mfrow=c(3,2) sau de mfcol=c(2,3) .

Pentru configurații mai complicate ale mai multor figuri, luați în considerare funcțiile layout() în layout{graphics} și split.screen() în screen{graphics} .

Drivere de dispozitiv

În calculul R , driverul de dispozitiv convertește instrucțiunile grafice (de exemplu, plot(x, y)) într-o formă pe care un anumit dispozitiv o poate accepta. Câteva drivere de dispozitiv comune sunt:

windows() Pentru utilizare pe Windows
quartz() Pentru utilizare pe MacOS X
X11() Pentru utilizare cu un sistem de ferestre X11 pe sisteme asemănătoare Unix
postscript() Pentru imprimarea pe imprimante PostScript sau crearea de grafice PostScript pdf() Pentru producerea unui fișier în format de document portabil (PDF)

png() Pentru producerea unei grafice de rețea portabile bitmap (PNG).
jpeg() Pentru producerea unui grup comun de experți fotografici (JPEG; bun pentru imagini)

Când rularea este încheiată, închideți driverul de dispozitiv după cum urmează:

```
> dev.off()
```

DIAGRAME POSTSCRIPT

Funcția grafică postscript() , de asemenea din pachetul grDevices , pornește driverul de dispozitiv grafic pentru producerea de grafică PostScript. Funcția are forma generală:

```
postscript(fișier = ifelse(un fișier, „Rplots.ps", „Rplot%03d.ps"),
           onefile, familie, titlu, fonturi, codificare, bg, fg,
           lățime, înălțime, orizontală, dimensiunea punctelor,
           hârtie, centru de pagină, print.it, comandă, model de culoare,
           useKerning, fillOddEven)
```

Trecând argumentul fișierului de către funcția driver de dispozitiv postscript() , se stochează graficul în format PostScript într-un fișier desemnat. Acea parcelă va fi în orientare peisaj dacă nu trece argumentul orizontal=FALSE . Dimensiunea graficului poate fi controlată folosind argumentele lățime și înălțime .

De exemplu, codul

```
> postscript("fișier.ps", orizontal=FALSE, înălțime=3, mărimea punctelor=8)
```

produce un fișier care conține cod PostScript pentru o cifră de 3 inci înălțime. Consultați documentația postscript {grDevices} pentru mai multe detalii.

DISPOZITIVE GRAFICE MULTIPLE

Atunci când este necesar să aveți mai multe dispozitive grafice deschise simultan, chiar și atunci când un singur dispozitiv poate accepta comenzi grafice la orice instanță dată (acela fiind dispozitivul **curent**), dispozitivele multiple formează o secvență numerotată cu nume care dă tipul de dispozitiv în orice poziție. . Fiecare apel nou către un dispozitiv extinde lista de dispozitive cu unul. Noul dispozitiv devine dispozitivul curent către care va fi apoi direcționată ieșirea grafică.

Această funcție R , dev() , de asemenea, în cadrul pachetului grDevices , este definită în una dintre următoarele forme:

```
dev.cur()
dev.list()
```

= returnează numerele și numele tuturor dispozitivelor active.

```
dev.next(care = dev.cur())
dev.prev(care = dev.cur())
```

= returnează numerele și numele următoarelor dispozitive active sau anterioare.

```
dev.off(care = dev.cur())
```

= termină dispozitivul grafic.

dev.set(care = dev.next())

= schimbă dispozitivul curent cu următorul, așa cum este specificat.

dev.new(...)

= returnează valoarea de returnare a dispozitivului deschis, de obicei invizibil NULL.

graphics.off()
= termină toate dispozitivele grafice din listă, cu excepția dispozitivului nul.

Aceste funcții pot fi utilizate cu următoarele argumente:

care = un număr întreg care specifică un număr de dispozitiv
… = argumente care trebuie transmise dispozitivului selectat

Întrebări de revizuire pentru Secțiunea 4.1

1. (a) Care sunt cele două sisteme de funcționalitate grafică R ?
   (b) Care sunt cele patru niveluri care constituie sistemul grafic R ?
2. (a) Care sunt cele trei grupuri de comenzi de trasare din R ? Descrieți pe scurt fiecare.
   (b) Sugerați o comandă R pentru instalarea unui pachet numit PACKAGENAME.
3. (a) Când pachetul PACKAGENAME a fost instalat, sugerați o comandă pentru a încărca toate funcțiile din acel pachet.
   (b) Dorim să obținem informații utile cu privire la o funcție numită FUNCȚIE. Sugerați o comandă pentru a obține documentația adecvată pentru acea funcție.
4. (a) Sugerați o funcție pentru a reprezenta grafic valorile unui vector arbitrar x.
   (b) Folosiți această funcție pentru a reprezenta graficul vector x = (1, 3, 5, 7, 9, 11, 13, 15, 17, 19).
5. *x* și *y* sunt vectori dați de *x* = (1, 3, 5, 7, 9) și *y* = (2, 4, 6, 8, 10). Construiți o matrice din două coloane, în care elementele lui *x* și *y* sunt elementele coloanei 1 și, respectiv, ale coloanei 2.
6. Sugerați o comandă pentru a reprezenta grafic elementele lui *x* și *y* (de la întrebarea 5) pentru a obține o diagramă de dispersie pe perechi a lui *x* și *y* ; apoi obțineți graficul de dispersie.
7. Pentru trasarea seturilor de date multivariate, descrieți graficele obținute de coplot().
8. Ce adaugă următoarele funcții de nivel scăzut la afișarea grafică a unei parcele deja finalizate?
   (a) text()
   (b) aline()
   (c) axă()
   (d) legendă()
9. (a) În grafica pachetului, ce parametri grafici pot fi setați de următorii parametri din funcția par() : "lty" , "lwd" și "col" ?
   (b) Când vă pregătiți pentru ieșiri de mai multe cifre, descrieți configurațiile diagramei obținute atunci când utilizați următorii parametri: mfcol , mfrow ,

oma și omi .

Exerciții pentru Secțiunea 4.1

1. Folosind R ca calculator, calculați răspunsurile la următoarele:
   (a) 1 + 2
   (b) 13 – 5
   (c) 17 × 29
   (d) 851/37
   (e) $(3{,}1416)^2$
   (f) π^2
   (g) e^{-3}
   (h) $\sqrt{(11^2 - 4 \times 3 \times 7)}$
   (i) jurnal$_{10}(1234567)$
   (j) păcatul$^2(30°)$
2. **Tensiunea arterială** este presiunea sângelui circulant împotriva pereților vaselor de sânge. Este măsurată ca parte a unei evaluări a sănătății unei persoane. Tensiunea arterială la adult este considerată normală la 120/80; primul număr este *presiunea sistolice* iar al doilea este ***presiunea diastolică*** .

   Presiunea sistolică este măsurată *în timpul* contracției ventriculului stâng al inimii, iar presiunea diastolică este măsurată ***după*** contracția inimii, în timp ce camerele inimii se umplu cu sânge.

   Următoarea este presiunea sistolica măsurată a pacientului A luată zilnic timp de 10 zile consecutive:

   145, 150, 135, 140, 160, 170, 138, 168, 155, 165

   (a) Introduceți aceste 10 citiri în variabila bpsistolic .
   (b) Utilizați funcția diff() pe această variabilă. Ce înseamnă rezultatele?
   (c) Utilizați comanda medie(bpsystolic) . Ce înseamnă rezultatele?
   (d) Utilizați comanda mean(diff(bpsystolic)) . Ce înseamnă rezultatele?
3. Folosind funcția boxplot() , introduceți cele 10 citiri ale tensiunii arteriale și obțineți un grafic al acestor 10 citiri (vezi Figura 4.11).

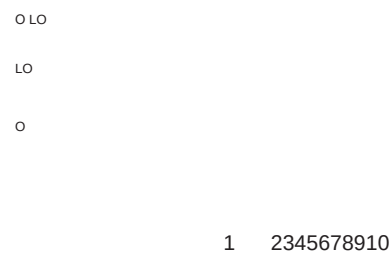


FIGURE 4.11 Plot of 10 blood pressure readings.

4. Patru anchete naționale succesive de sănătate, pentru populația de subiecți de sex masculin în vârstă de 20 de ani, au arătat că cantitatea medie de calorii zilnice a

fost:

2450, 2439, 2866, 2618.

Procentul de calorii din grăsimi a fost de 37,0%, 36,2%, 34,0%, 32,1%.
Procentul de calorii din carbohidrați a fost de 43,1%, 42,2%, 50,0%, 48,1%.

(a) Crește sau scade numărul mediu de calorii în grăsimi?
(b) Este acest rezultat în concordanță cu informațiile că în aceeași perioadă de timp, prevalența obezității în țară a crescut de la 14,5% la 30,9%?

5. Pentru datele din exercițiul 4, utilizați funcția boxplot() pentru a nota cele trei comenzi pentru a obține grafice ale nivelurilor relative ale:
   (a) calorii
   (b) procentul de calorii din grăsimi
   (c) procentul de calorii din carbohidrați

```
> boxplot(2450, 2439, 2866, 2618)
> boxplot(37.0, 36.2, 34.0, 32.1)
> boxplot(43.1, 42.2, 50.0, 48.1)
```

Rezultatele sunt prezentate în figurile 4.12, 4.13 și 4.14 .

6. Următoarele sunt câteva date privind ratele accidentelor pe grupe de vârstă (Dalgaard, 2002) . Grupele de vârstă sunt 0–4, 5–9, 10–15, 16, 17, 18–19, 20–24, 25–59 și 60–79 de ani. Datele înregistrate sunt rezumate după cum urmează:

```
> grup.midage<-c(2,5, 7,5, 13, 16,5, 17,5, 19, 22,5, 44,5, 70-5)
> accidente <- c(28, 46, 58, 20, 31, 64,149, 316, 103)
```

Combinați acești doi parametri după cum urmează:

```
> age.acc <- rep(group.midage, accidente)
```

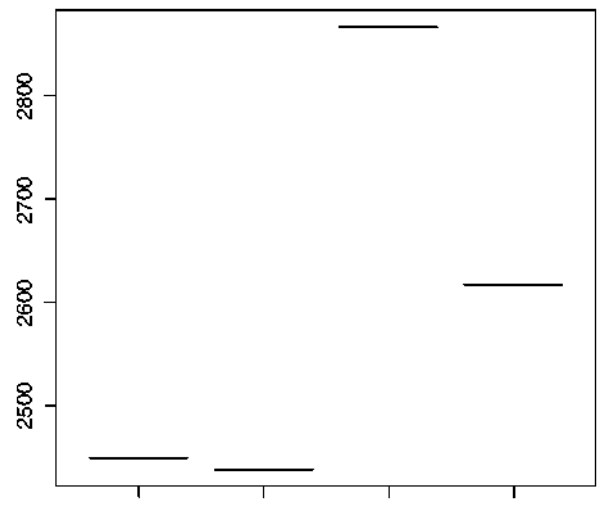


FIGURE 4.12 Total calories for four tests.

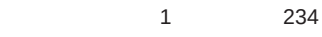


FIGURA 4.13 Procentul de calorii din grăsimi pentru patru teste.



FIGURA 4.14 Procentul de calorii din carbohidrați pentru patru teste.

Notă : Funcția rep(x) reproduce valorile din *x* .

Acum, definiți punctele de întrerupere după cum urmează:

```
> punct de întrerupere <- c(0, 5, 10, 16, 17, 18, 20, 25, age.acc
60, 80)
```

Trasează o histogramă a parametrului age.acc , pentru a afișa distribuția pe grupe de vârstă, după cum urmează:

```
> hist(age.acc, breaks=breakpoint)
```

și obțineți histograma factorului de vârstă.acc (Figura 4.15).

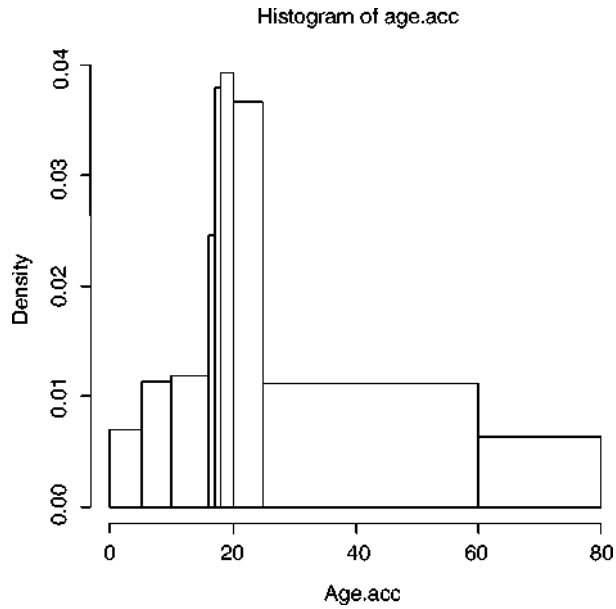


FIGURA 4.15 Histograma parametrului vârstă–acc.

7. În pachetul HSAUR este un set de date apă , care este o înregistrare a 61 de orașe din Marea Britanie cu informații despre distribuțiile marginale ale durității apei (concentrația de calciu) și mortalitatea.

   (a) Accesați acest set de date prin:
   > date("apă", pachet="HSAUR")

   (b) Examinați setul de date după:
   > apa

   (c) Obține un grafic de dispersie a mortalității în funcție de duritate prin:
   > parcela (date = apa)

   (d) Trasează linia de regresie liniară a mortalității în funcție de duritate după:
   > abline(lm(mortalitate ~ duritate, date = apă))

   (e) Adăugați un tabel de legendă în colțul din dreapta sus al graficului prin:
   > legenda("topright", legenda = niveluri(apa$locatie),
   + pch=c(1, 2), bty= "n")

   (f) Afișați o histogramă a apei în funcție de duritate prin:
   > hist(duritatea apei)

   (g) Afișați o diagramă cu casetă a apei în raport cu mortalitatea prin:
   > boxplot(apă$mortalitate)

   Rezultatele sunt prezentate în figurile 4.16–4.20.

8. Afișarea datelor multivariate.

   Într-un studiu danez privind efectul screening-ului pentru cancerul de sân (Dalgaard, 2002;

   Olsen et al., 2005) , au fost colectate patru grupuri sau cohorte:

   (i) Grupul de studiu, format din populația de femei din intervalul de vârstă corespunzător din Copenhaga și Frederiksberg, după introducerea screening-ului mamografic de rutină

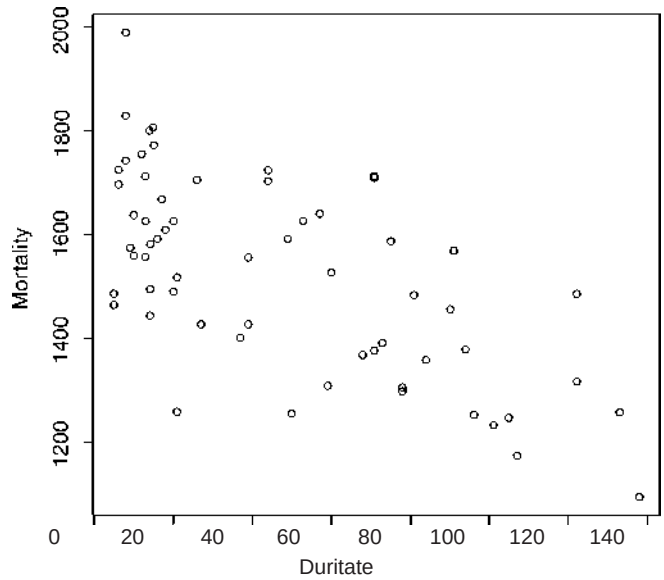


FIGURA 4.16 Mortalitatea vs duritatea apei.

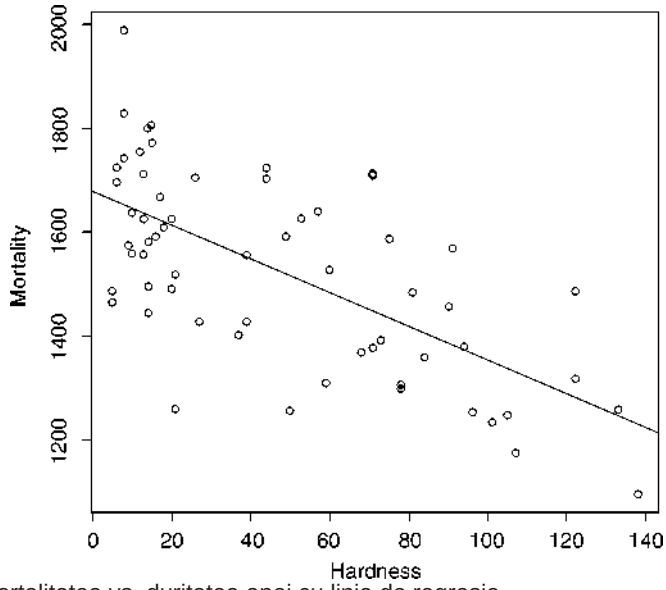


FIGURA 4.17 Mortalitatea vs. duritatea apei cu linia de regresie.

(ii) Grupul național de control, format din populația din părțile Danemarcei în care screeningul mamografic de rutină *nu a fost* disponibil

Aceste două grupuri au fost colectate în 1991–2001.

(iii) Grupul de control istoric

(iv) Grupul istoric de control național

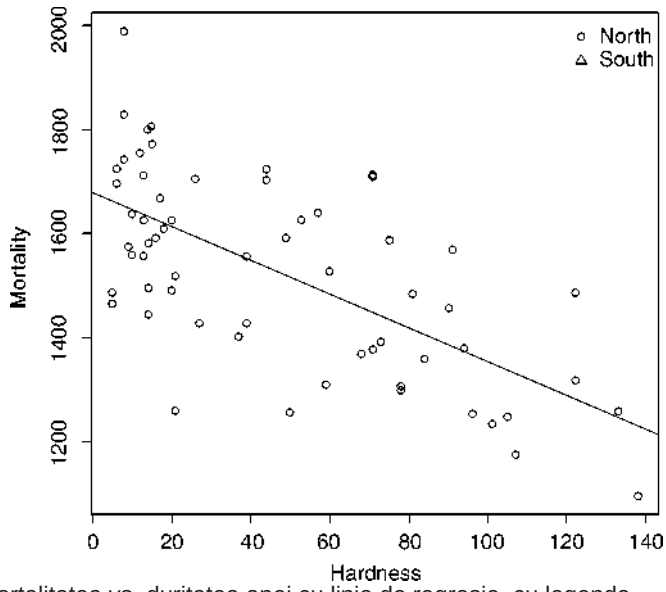


FIGURA 4.18 Mortalitatea vs. duritatea apei cu linie de regresie, cu legenda.

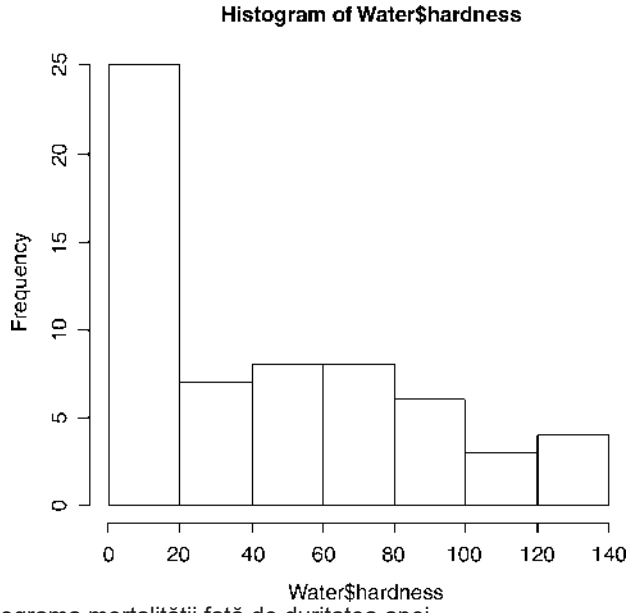


FIGURA 4.19 Histograma mortalității față de duritatea apei.

Acestea din urmă sunt cohorte similare cu 10 ani mai devreme, înainte de introducerea screening-ului la Copenhaga și Frederiksberg. Grupul de studiu cuprinde întreaga populație, nu doar cei care acceptă invitația de a fi examinați.

(v) Examinați setul de date, folosind următorul segment de cod R :

```
> install.packages ("ISwR")
> bibliotecă ("ISwR")
```

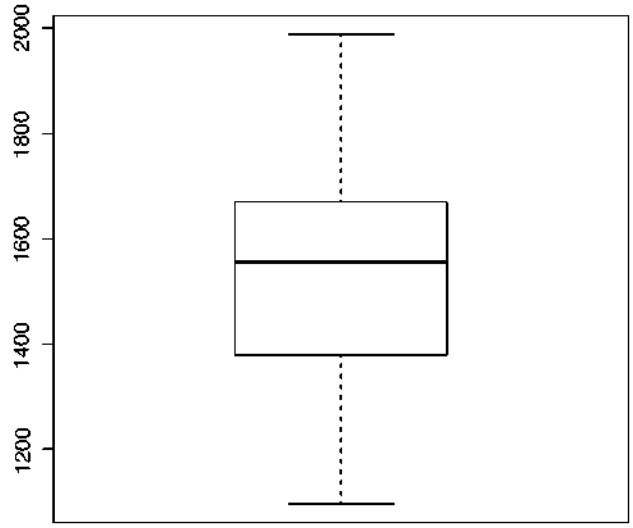


FIGURA 4.20 Boxplot a datelor.

```
> date (bcmort)
> bcmort
```

(vi) Afișează setul de date după:

```
> complot(bcmort)
```

(vii) Eliminați golurile folosind funcția pereche() :

```
> par(mex=1)
> perechi (bcmort, gap=0, cex.labels=2,0)
```

Rezultatele sunt prezentate în figurile 4.21 și 4.22.

9. Afișarea mai multor date multivariate.

Un studiu de sănătate publică a investigat efectul greutății corporale asupra ratei metabolice de repaus (rmr) la femei (Dalgaard, 2002 ; Altman, 1991).

Cadrul de date rmr are 44 de rânduri și 2 coloane, care conține datele rmr și greutatea corporală pentru 44 de femei. Cele două coloane sunt:

greutate.corp Un vector numeric, greutate corporală (kg) rata metabolică Un vector numeric, rata metabolică (kcal/24 ore)

(a) Examinați setul de date, folosind următorul segment de cod R :

```
> install.packages ("ISwR")
> bibliotecă ("ISwR")
> date(rmr)
> rmr
```

(b) Afișează setul de date după:

```
> complot(rmr)
```

Rezultatul este prezentat în Figura 4.23.

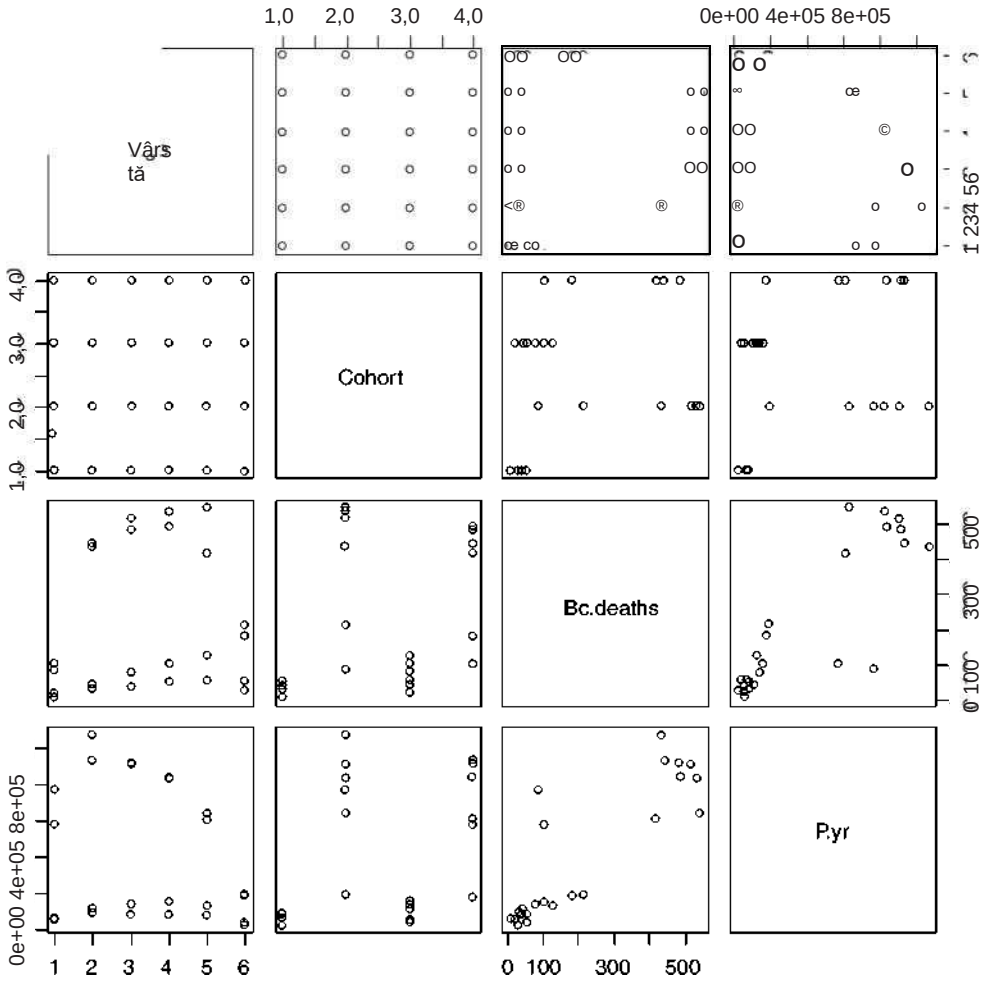


FIGURA 4.21 diagrame de
perechi de date.

(c) Executați următorul complot:
> plot(metabolic.rate~body.weight,data=rmr)

Rezultatul este prezentat în Figura 4.24 .

(d) Observați vreo diferență între acest complot și ultimul?
(e) Adăugați o linie de regresie liniară pe afișaj folosind:
> abline(lm(metabolic.rate ~ body.weight, data = rmr))

Rezultatul este prezentat în Figura 4.25 .

10. O procedură pas cu pas pentru afișarea unei parcele cu etichetare (Murrell, 2006) .

(a) Pregătește-te utilizând:
> plot.new()
(b) Configurați o fereastră utilizând:
> plot.window(interval(presiune$temperatura),
+ interval (presiune$temperatura))

FIGURE 4.23 pairs plots of data.

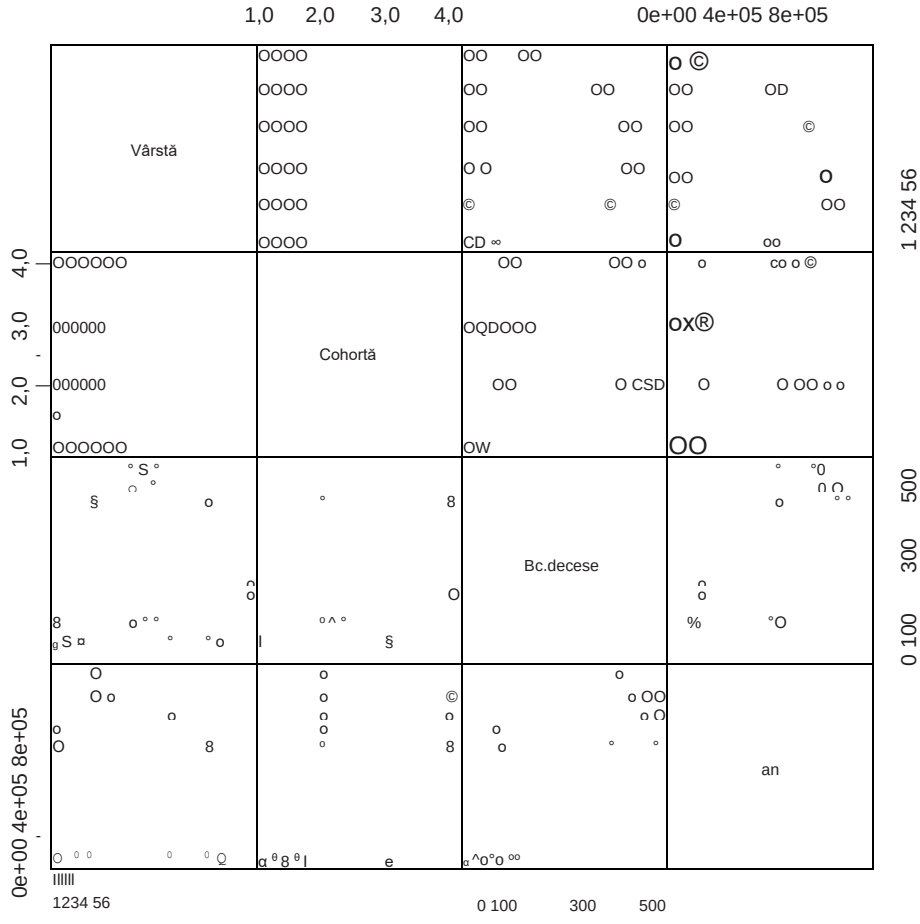
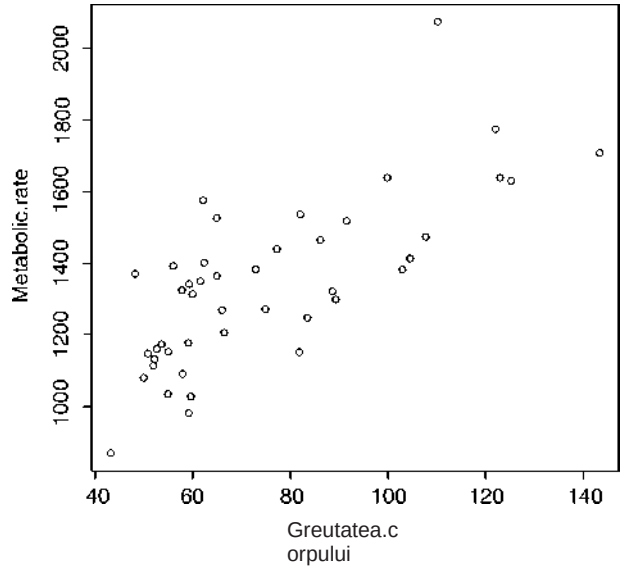


FIGURA 4.22 perechi de diagrame de date, eliminând golurile dintre diagramele individuale.



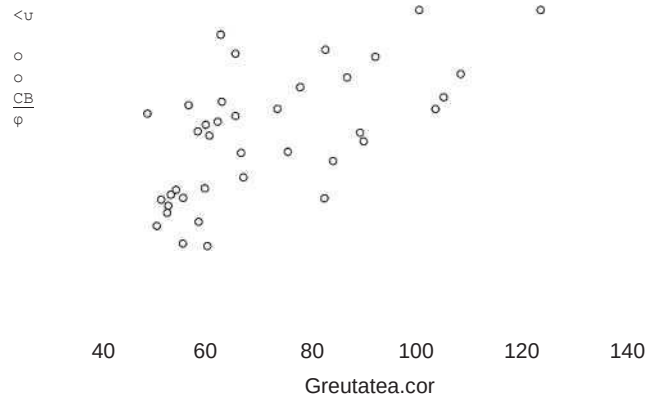


FIGURA 4.24 Graficul specific al setului de date.

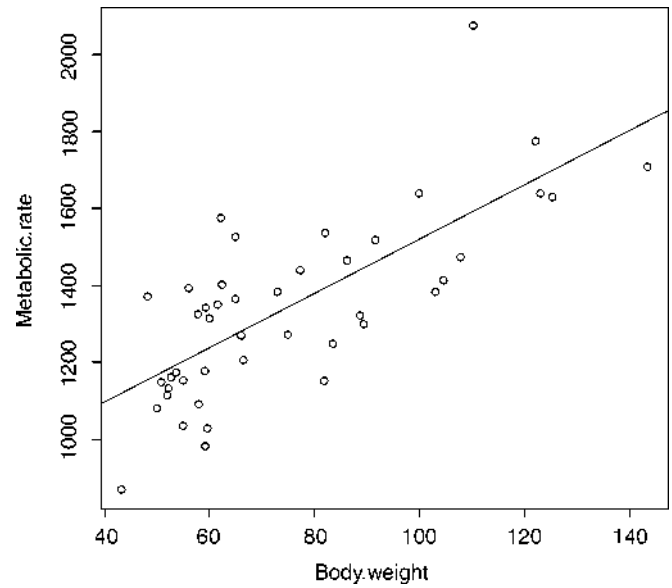


FIGURA 4.25 Graficul specific al setului de date, cu linia de regresie.

(c) Trasează datele despre presiune în funcție de temperatură utilizând:

```
> plot.xy(pressure, type="p")
> # Ieșire: Figura 4.26 .
```

(d) Puneți un cadru dreptunghiular peste afișaj folosind:

```
> cutie()
> # Ieșire: Figura 4.27 .
```

FIGURE 4.23 pairs plots of data.

FIGURA 4.26 > plot.xy(pressure, type="p").

FIGURA 4.27 Adăugarea > box().

(e) Adăugați axa orizontală (temperatură) utilizând:
> axa(1)
> # *Ieșire:* Figura 4.28 .

(f) Adăugați cealaltă axă, axa verticală (de presiune), utilizând:
> axa(2)
> # *Ieșire:* Figura 4.29 .

(g) În cele din urmă, etichetați diagrama centrată la poziție (100 de unități orizontală, 250 de unități verticală) utilizând:

```
> text(100, 250, "Presiune (mm Hg)\nversus\nTemperatura
+ (Centigrad)")
> # Ieșire: Figura 4.30 .
```

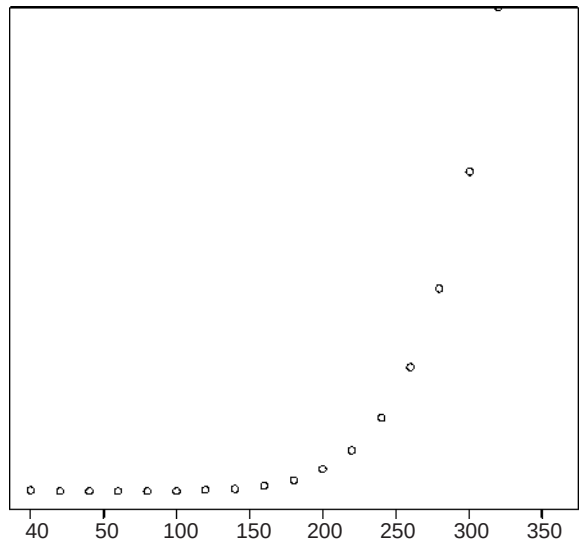


FIGURA 4.28 Adăugarea > axa(1).

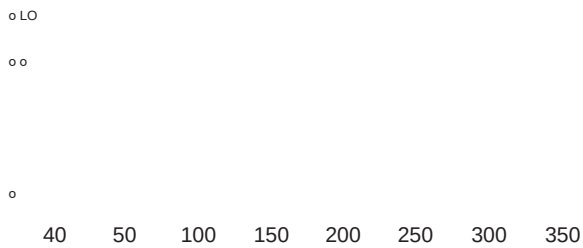


FIGURA 4.29 Adăugarea > axa(2).

Presiune (mmHg)
față de
temperatură

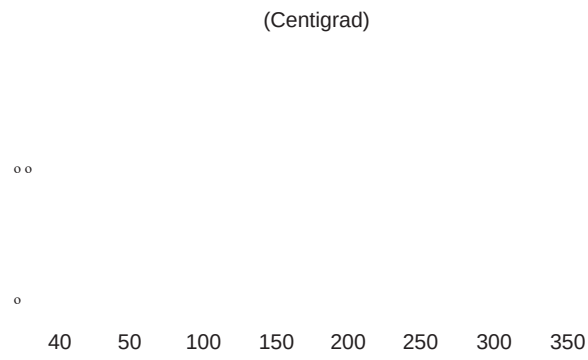


FIGURA 4.30 Adăugarea > text(100, 250, „Presiune (mm Hg)\nversus\nTemperatura (Centigrad)").

4.2 GRAFICA GRILĂ

În mediul R pentru calcul statistic și grafică, sistemul **Grid Graph ics** este un pachet suplimentar. Oferă un set de funcții grafice, cu o flexibilitate substanțială, care acceptă afișarea grafică. Câteva alte pachete R folosesc, de asemenea, grafica grilă (Murrell, 2006), inclusiv:

- cu zăbrele al lui Deepayan Sarkar , distribuit cu R
- Hmisc și design al lui Frank Harrell
- RGrace a lui M. Kondrin
- GridBase lui Paul Murrell (disponibil de la CRAN) și gridSVG (disponibil de pe pagina de start a lui Murrell)

Sistemul de grafică grilă face acum parte din distribuția de bază R.

Această secțiune introduce grafica grilă în termenii pachetului seminal CRAN lattice (Sarkar, 2011b) și tratatul definitiv al lui Murrell (2006) .

Pachetul cu zăbrele : grafică trellis

Pachetul Lattice (http://r-forge.r-project.org/projects/lattice) este un sistem de vizualizare a datelor puternic, elegant, la nivel HYPERLINK "http://r-forge.r-project.org/projects/lattice"înalt, cu accent pe datele multivariate, care este suficient pentru nevoile grafice tipice. De asemenea, este suficient de flexibil pentru a face față majorității cerințelor nestandard. (Strict vorbind, grafica cu zăbrele produce un obiect din clasa „spalier", care conține o descriere a plotului, iar funcția print() desenează plotul.)

Ca o introducere în grafica grilă, în această secțiune, considerăm una dintre cele mai cunoscute funcții grafice latice , xyplot() , ca metodă de trasare a seriilor de timp. Alte funcții de grilă utile sunt, de asemenea, prezentate în exemplele următoare. Funcția xyplot () se ocupă de trasarea serii de timp, inclusiv de diagrame de tăiere și stivuire și permite suprapunerea, juxtapunerea și stilizarea diferitelor serii temporale. Luați în considerare următorul segment de cod (Sarkar, 2011a) :

```
> bibliotecă (zăbrele)
> install.packages(„grafică")
> bibliotecă ("grafică")
```

rețeaua pachetului sunt selectate și rulate în mediul R. Sunt afișate graficele rezultate .

> **Exemplul 4.8:** Grafică grilă cu zăbrele

xyplot.ts Metode de trasare a seriilor de timp (Sarkar, 2011a)

```
> ### Exemplu cu date mai simple, câteva puncte de date
> set.seed(1)
> z <- ts(cbind(a = 1:5, b = 11:15, c = 21:25) + rnorm(5))
> xyplot(z, ecrane = 1)
> # Ieșire: > xyplot(z, ecrane = listă(a = "primar (a)", "altele (b & c)"),
+ tip = listă (a = c ("p", "h"), b = c ("p", "s"), "o"),
+ pch = listă (a = 2, c = 3), auto.key = listă (tip = "o"))
> # Ieșire: figurile 4.31 și 4.32 .
```

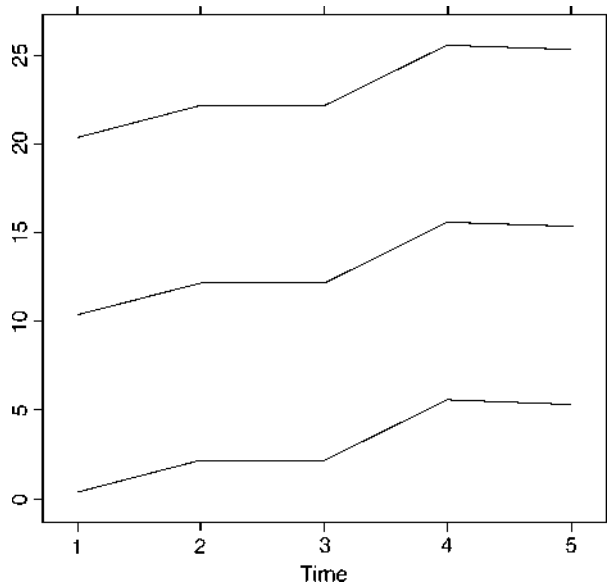


FIGURA 4.31 Grafică grilă cu zăbrele : xyplot() .

abc

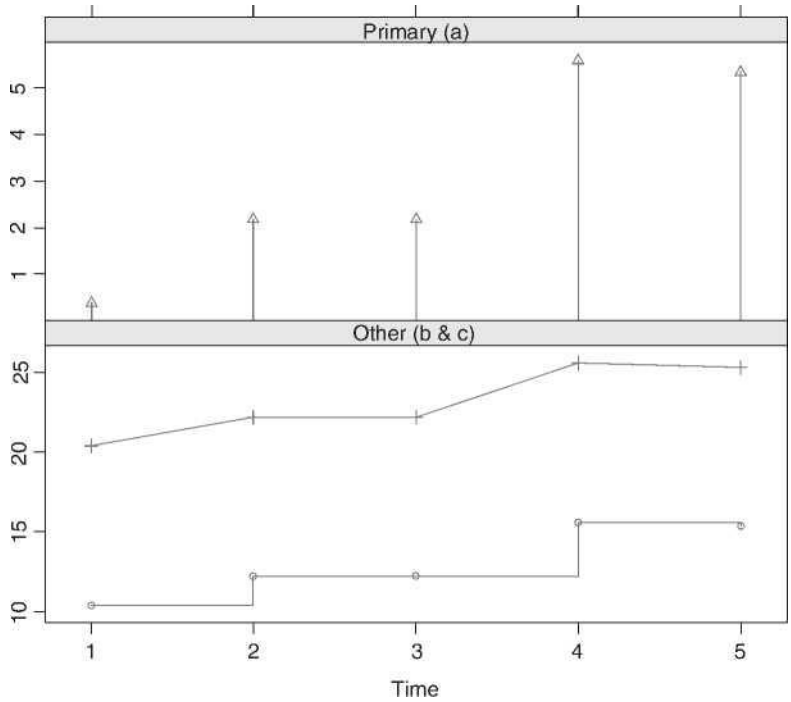


FIGURA 4.32 Grafică grilă : xyplot() , separând un set de date în două grupuri.

> **Exemplul 4.9:** Norul grafic al grilei lattice (x, date, ...)

> ?cloud *# Este prezentată următoarea documentaţie* [18]: B_07_cloud {zăbrele}

Graficul de dispersie 3D și graficul de suprafață cu cadru fir

Descriere:
Funcții generice pentru a desena diagrame și suprafețe de dispersie tridimensionale (3D). Metodele formulei fac cea mai mare parte a muncii efective.

Utilizare:
nor (x, date, ...)
wireframe (x, date, ...)

Detalii:
Aceste funcții produc diagrame 3D în fiecare panou (atâta timp cât sunt utilizate funcțiile implicite ale panoului). Orientarea se obține după cum urmează: Datele sunt scalate pentru a se încadra într-o casetă de delimitare care este conținută în cubul [0,5, 0,5] (și chiar mai mici pentru valorile non-implicite ale aspectului). Direcția de vizualizare este dată de o secvență de rotații specificată

prin argumentul ecranului , pornind de la axa *z* pozitivă . Punctul de vizualizare (camera) este situat la o distanta de 1/distanta de la origine. Dacă perspectiva=FALSE, distanța

[18] Site CRAN pentru grid(){graphics}:
http://127.0.0.1:19796/library/graphics/html/grid.html .

este setată la 0 (adică punctul de vizualizare este la o distanță infinită).

cloud desenează un grafic de dispersie 3D, în timp ce wireframe desenează o suprafață 3D (de obicei evaluată pe o grilă). Mai multe suprafețe pot fi desenate prin wireframe folosind argumentul grupurilor (deși acest lucru este de utilitate limitată deoarece afișajul este incorect atunci când suprafețele se intersectează). Specificarea grupurilor cu nor are ca rezultat un efect de tip panel.superpose (prin panel.3dscatter).

wireframe poate reda opțional suprafața ca fiind iluminată de o sursă de lumină (fără umbre, totuși). Detalii pot fi găsite în pagina de ajutor pentru panel.3dwire . Rețineți că, deși argumentele care le controlează sunt de fapt argumente pentru funcția panoului, ele pot fi aplicate direct în cloud și wireframe .

Pentru diagramele cu un singur panou, wireframe poate reprezenta și suprafețe 3D parametrizate [adică, funcții de forma f(u,v) = (x(u,v), y(u,v), z(u,v)], unde valorile (u,v) se află pe un dreptunghi Cel mai simplu exemplu de acest tip de suprafață este o sferă parametrizată prin latitudine și longitudine wireframe cu o formulă x de forma z~x*y , unde x , y , și z sunt toate matrice de aceeași dimensiune, reprezentând valorile lui x(u,v), y(u,v) și z(u,v) evaluat pe o grilă dreptunghiulară discretă [valorile reale ale lui (u,v) sunt irelevante].

Când se utilizează această caracteristică, înălțimile folosite pentru a calcula culorile drapei sau culorile de umbrire nu mai sunt valorile z , ci distanțele (x,y,z) de la origine.

Rețineți că această caracteristică nu funcționează cu grupuri , indice , subseturi sau alte asemenea funcții. De asemenea, variabilele de condiționare nu sunt acceptate în acest caz.

Algoritmul pentru identificarea marginilor casetei de delimitare sunt „în spatele" punctelor nu funcționează în unele situații extreme. De asemenea, panel.cloud încearcă automat să descopere locația optimă a săgeților și a etichetelor axelor, dar poate eșua ocazional (mai ales când vizualizarea este de „sub" date). Acest lucru poate fi controlat manual de argumentul scpos din panel.cloud .

Acestea și toate celelalte funcții Trellis de nivel înalt au câteva alte argumente în comun. Acestea sunt documentate pe larg doar în pagina de ajutor pentru xyplot , care ar trebui consultată pentru a afla mai multe detalii despre utilizare.

```
> cloud(Sepal.Length ~ Petal.Length * Petal.Width | Species, + data=iris,
screen=list(x=-90, y=70), distance =0.4, zoom =0.6) > # ieșire: Figura 4.33 .
```

CONTROLUL PARCELELOR DE ZARELE

Pentru a controla culoarea, dimensiunea fontului textului, tipurile/lățimile de linii ale afișajului grafic și așa mai departe într-o diagramă lattice , pot fi utilizați parametrii grafici. Liste mari de grupuri de parametri se află în fișierele de documentație ale fiecărei funcții latice și fiecare grup de parametri constă dintr-o listă de setări ale parametrilor. Pentru fiecare funcție, informațiile de referință sunt disponibile online, împreună cu ruta de ajutor (funcție) din mediul

R.

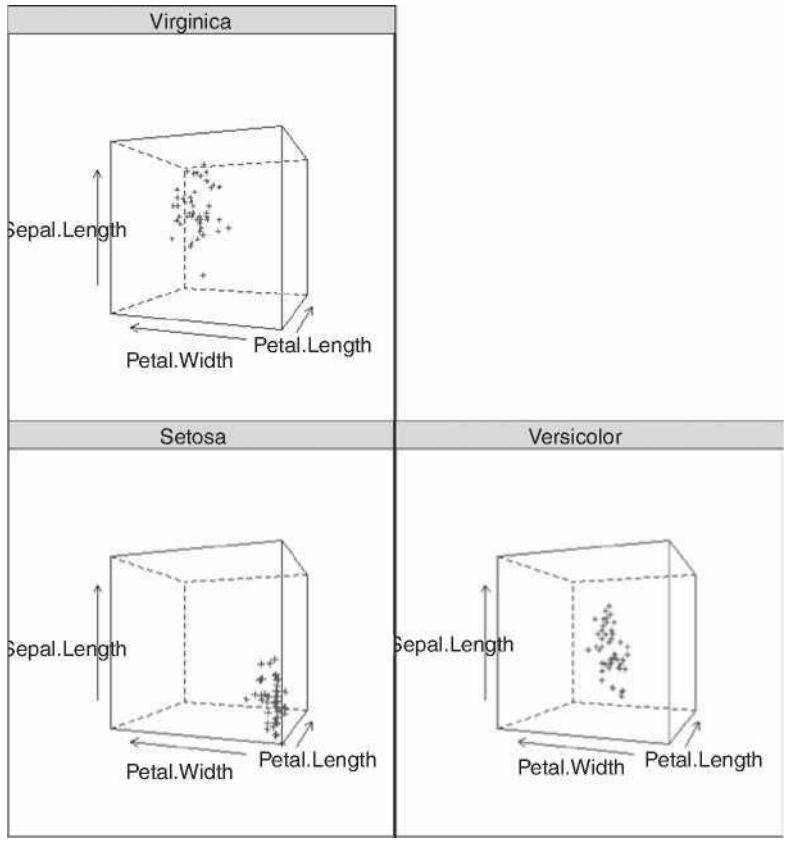


FIGURA 4.33 Norul grafic al grilei lattice (x, date, ...).

> *Controlul dimensiunii fontului textului:*

> **Exemplul 4.10:** Pentru a controla dimensiunea fontului textului graficului latice din Exemplul 4.9

Aici, se poate folosi următorul segment de cod pentru a controla dimensiunea fontului:

```
> fontsize <- trellis.par.get("fontsize")
> fontsize$text <- 10
> trellis.par.set(„dimensiunea fontului", dimensiunea fontului)
>
> nor(Sepal.Lungime ~ Petal.Length * Petal.Width | Specie,
+ date=iris, ecran=listă (x=-90, y=70), distanță = 0,4, zoom = 0,6)
> # Ieșire: Figura 4.34 .
```

> *Controlul culorii/tipului/lățimii liniei:* Consultați plot.lone .
> *Controlul simbolurilor de date, dimensiunea, forma și culoarea:* Consultați setările plot.symbol și pch , precum și setările fontsize și strip.background .
> Valoarea curentă a setării parametrilor grafici poate fi obținută folosind funcțiile trellis.par.get() .
> Setările pentru dimensiunea fontului pot fi specificate folosind funcția trellis.par.set() sau folosind argumentul par.settings dintr-o comandă de plotare.

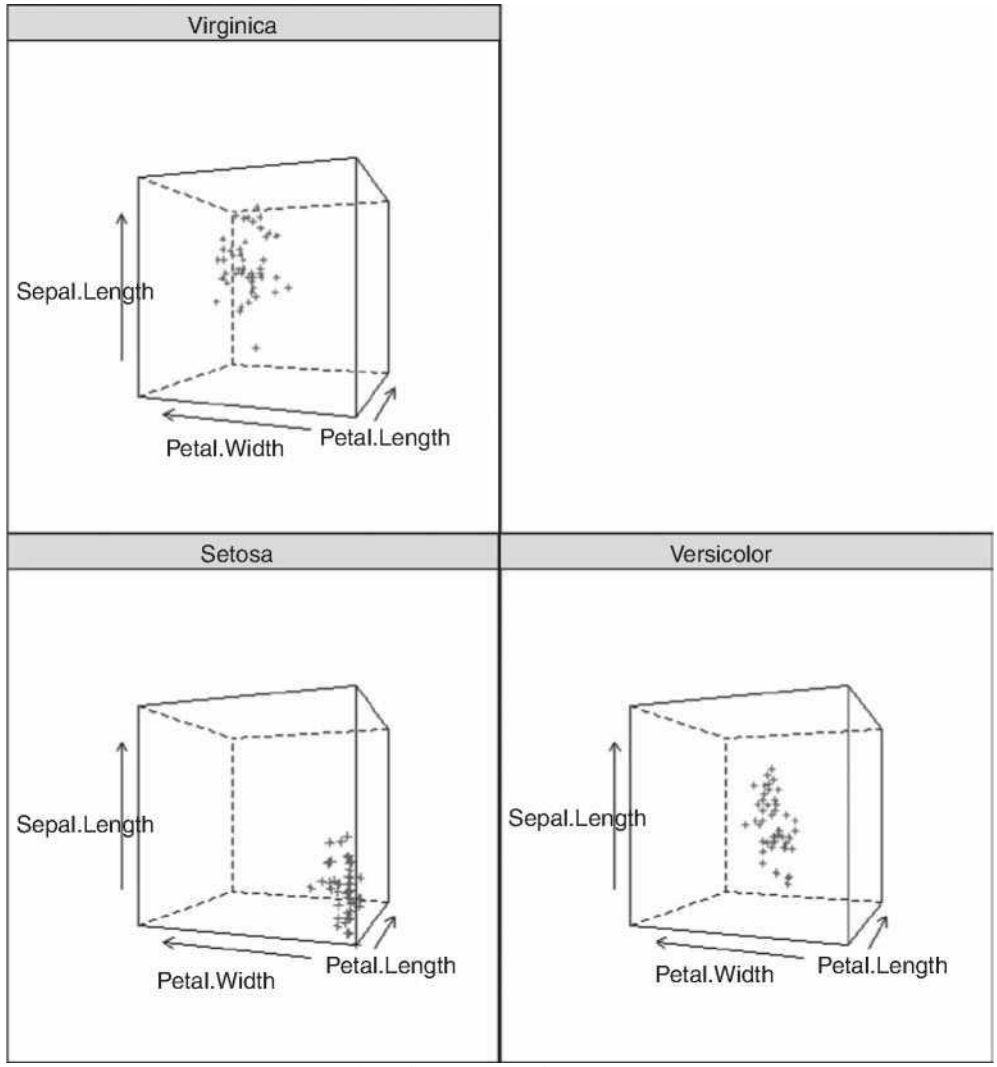


FIGURA 4.34 Norul grafic al grilei lattice (x, date, ...) cu dimensiunea fontului modificat folosind > fontsize$text.

AMENAJARE PARCELE DE ZARELE

Două tipuri de aranjamente trebuie luate în considerare atunci când se realizează diagrame cu zăbrele :

Tip I: În cadrul unei singure parcele, aranjarea benzilor și panourilor
Tip II: Pe o pagină dată, aranjarea mai multor parcele complete împreună

Pentru Tipul I, pot fi specificate două argumente: argumentul aspect și argumentul aspect .

1. Argumentul aspectului este format din până la trei valori; primele două indică numărul de rânduri și coloane de panouri de pe fiecare pagină, iar al treilea indică numărul de pagini.
2. Argumentul aspect specifică raportul de aspect (înălțimea împărțită la lățime)

panoului. Valoarea implicită este umplere , ceea ce face ca panourile să se extindă pentru a umple cât mai mult spațiu posibil. aspect=1 forțează toate panourile să fie pătrate, deoarece un pătrat are un raport de aspect de 1,0.

Modelul Grid pentru R Graphics (Murrell, 2006) [3]

După cum sa menționat mai devreme, sistemul de grafică Grid este un pachet suplimentar din mediul R. Sistemul de grilă este format din caracteristici de bază precum funcții pentru desenarea obiectelor geometrice (punct, linii, triunghiuri, dreptunghiuri etc.); texte; și concepte precum machete, ferestre și unități, care permit dimensionarea și localizarea ieșirilor după cum este necesar. Sistemul de grilă poate fi încărcat în R după cum urmează:

```
> bibliotecă (grilă)
```

cu documentație online suplimentară accesibilă prin funcțiile help() și vignette() .

```
> ?grid # Pentru a examina grila de funcții din grila pachetului grafic , adaugă o
```

grilă dreptunghiulară nx cu ny la o diagramă existentă.

Utilizare:

```
> grid(nx = NULL, ny = nx, col = „gri deschis", lty = „punct",
+ lwd = par("lwd"), echilogs = TRUE)
```

Argumente:

nx , ny Numărul de celule ale grilei în direcțiile x și y . Când NULL , ca implicit, grila se aliniază cu marcajele de pe axa *implicită corespunzătoare* (adică, marcajele calculate de axTicks). Când NA , nu sunt trasate linii de grilă în direcția corespunzătoare.

Col Caracter sau (întreg) numeric; culoarea liniilor grilei.

Lty Caracter sau (întreg) numeric; tipul de linie al liniilor de grilă.

Lwd Numeric nenegativ care dă lățimea liniilor liniilor grilei.

echiloguri logice; folosit numai atunci când coordonatele *jurnalului* și alinierea cu marcajele axei sunt active. Setarea echilogs = FALSE în acest caz oferă linii de grilă aliniate cu bifă *non-echidistante* .

> **Exemplul 4.11:** O demonstrație a sistemului de grafică grilă

```
> complot (1:3)
> grid(NA, 5, lwd = 2) # grilă numai în direcția y
> # Ieșire: Figura 4.35 .
> ## poate schimba numărul dorit de semne de bifare:
> ## par(lab=c(mx,my,7))
> op <- par(mfcol = 1:2)
```

```
> cu (iris,
+       {
+       plot(Sepal.Length, Sepal.Width, col = as.integer(Specie),
+ xlim = c(4, 8), ylim = c(2, 4.5), panel.first = grid(),
+ main = "cu (iris, plot( ..... , panel.first = grid(), ..) )")
+ plot(Sepal.Length, Sepal.Width, col = as.integer(Specie),
+ panel.first = grid(3, lty=1,lwd=2),
+ main = "... panel.first = grid(3, lty=1,lwd=2), ..")
+       }
+       ) # Ieșire: Figura 4.36 .
>   # La sfârșitul trasării, resetați la setările anterioare:
>   par (op)
```

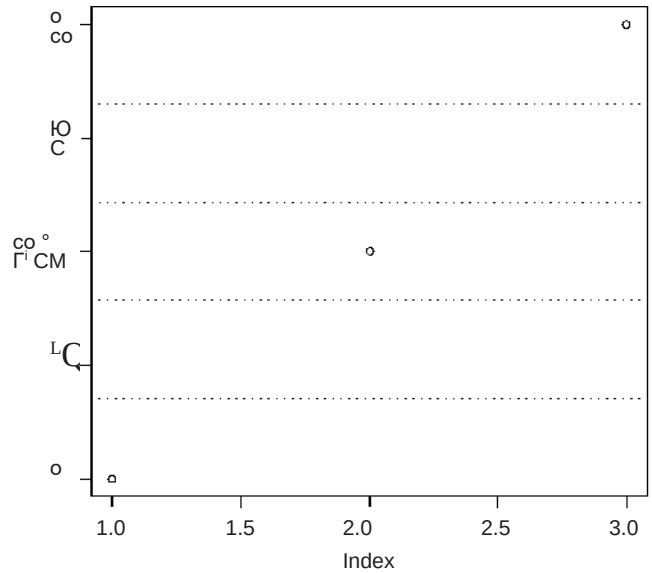


FIGURE 4.35 Grid Graphics System plot(1:3) with grid only in the *y*-direction.

> Controlul ieșirii :

Există un set de funcții grilă pentru a produce rezultate grafice de bază, cum ar fi puncte, linii, dreptunghiuri, cercuri și text. Aceste funcții au forma grilă.*(). Pentru fiecare dintre ele, există o funcție Grob() corespunzătoare care creează un obiect care conține o descriere a ieșirii grafice primitive, dar nu desenează nimic.

Setul complet al acestor funcții este listat în Tabelul 4.1 . [3]

Note:

1. În cele mai multe cazuri, primul argument al fiecăreia dintre aceste funcții este un set de locații și dimensiuni pentru obiectul grafic de desenat. De exemplu, grid.rect() are argumente x , y , lățime și înălțime , specificând locațiile și dimensiunile dreptunghiurilor de desenat. O excepție este funcția grid.text() , care necesită ca textul să fie desenat ca prim argument.
2. Pot fi produse mai multe primitive atunci când sunt specificate mai multe locații și dimensiuni.

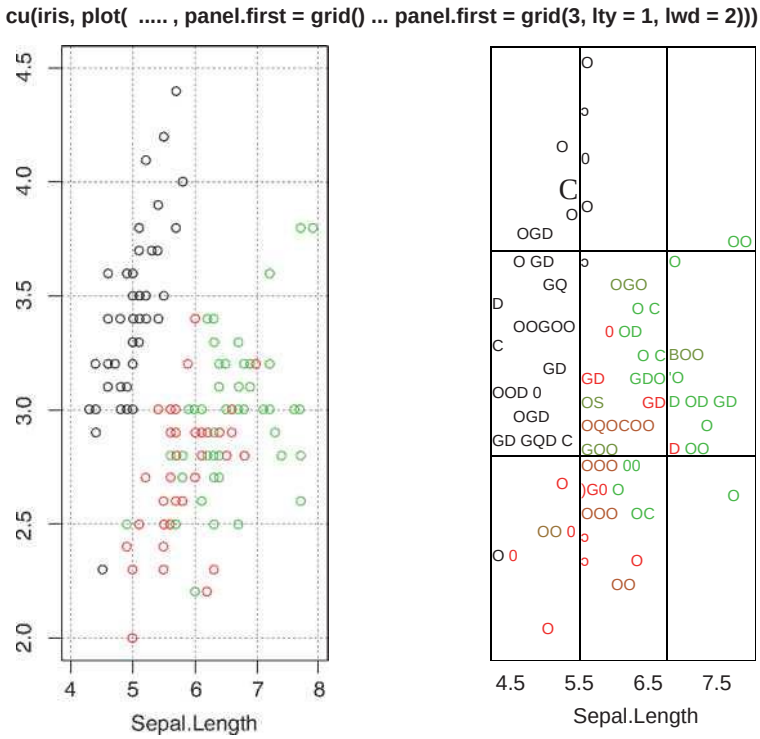


FIGURA 4.36 Sistemul grafic al grilei plot() cu linii, titluri, etichete și date specificate ale grilei puncte în culori diferite.

TABELUL 4.1 Primitive grafice în Grid

Acesta este setul complet de funcții de nivel scăzut care produc rezultate grafice. Pentru fiecare funcție care produce ieșire grafică (coloana din stânga), există o funcție corespunzătoare care returnează un obiect grafic care conține o descriere a ieșirii grafice în loc să producă ieșire grafică în sine (coloana din dreapta). Ultimul set de funcții este descris mai târziu în acest capitol.

FUNCȚIE DE PRODUCERE IEȘII	DESCRIERE	FUNCȚIA DE PRODUCERE OBIECTUL
grid.move.to()	Setați locația curentă.	moveToGrob()
grid.line.to()	Desenați o linie de la locația curentă la o nouă locație și resetați locația curentă.	lineToGrob()
grid.lines()	Desenați o singură linie prin mai multe locații în secvență.	linesGrob()
grid.segments()	Desenați mai multe linii între perechi de locații.	segmentsGrob()
grid.rect()	Desenați dreptunghiuri date locații și dimensiuni.	rectGrob()

(*continuare*

)

TABELUL 4.1 Primitive grafice în Grid (*continuare*)

FUNCȚIE DE PRODUCERE IEȘII	DESCRIERE	FUNCȚIA DE PRODUCERE OBIECTUL
grid.cercle()	Desenați cercuri cu locații și raze date.	circleGrob()
grid.poligon()	Desenați poligoane date vârfuri.	polygonGrob()
grid.text()	Desenați textul dat șiruri, locații și rotații.	textGrob()
grid.sageti()	Desenați săgeți la fiecare capăt al liniilor date locații	arrowsGrob() sau un obiect care descrie linii.
grid.points()	Desenați simboluri de date în locații date.	puncteGrob()
grid.xaxis()	Desenați axa *x* .	xaxisGrob()
grid.yaxis()	Desenați axa *y* .	yaxisGrob().

Obiecte grafice grilă

În grafica grilă, un **grob** este un *obiect grafic* . Un grob poate fi folosit pentru a edita interactiv un scenariu produs de grid . Deoarece rețeaua este construită pe grilă , această abordare permite editarea interactivă a unui diagramă reticulat .

În următorul segment de cod, care are 10 ieșiri grafice, observați funcția grob barbedGrob() , descrisă în pachetul CRAN gridExtra (Baptiste, 2012) .

```
> set.seed(1234)
> grid.barbed(name="test") # Ieșire: Figura 4.37 .
> grid.edit("test", gp=gpar(fill="albastru", lwd=3))
> # Ieșire: Figura 4.38 .
> grid.edit("test::points", pch=22) # Ieșire: Figura 4.39 .
> grid.newpage()
> g <
> barbedGrob(size=unit(1:5, "char"), only=FALSE,
+ gp=gpar(col="roșu", lex=3, umplere="albastru", alfa=0,5, pch=3))
>
> pushViewport(vp=port view(lățime=1, înălțime=1))
> grid.rect(gp=gpar(fill="thistle2")) # Ieșire: Figura 4.40 .
> grid.grill(gp=gpar(col="lavenderblush1",lwd=3,lty=3))
> # Ieșire: Figura 4.41 .
> grid.draw(g) # Ieșire: Figura 4.42 .
> x <- c(0,2, 0,7)
> y <- x
> dev.new(width=3, height=7) # Ieșire: Figura 4.43.
> grid.newpage() # Ieșire: Figura 4.44.
> grid.draw(g) # Ieșire: Figura 4.45.
> grid.points(x, y, pch=3) # Ieșire: Figura 4.46.
```

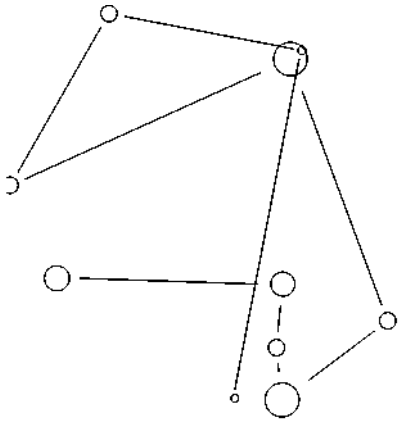


FIGURE 4.37 grid.barbed() Output.

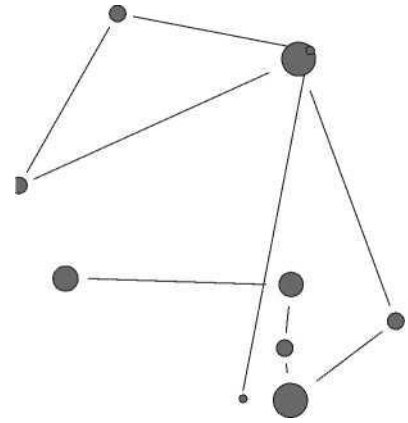


FIGURA 4.38 Ieșire grid.edit() .

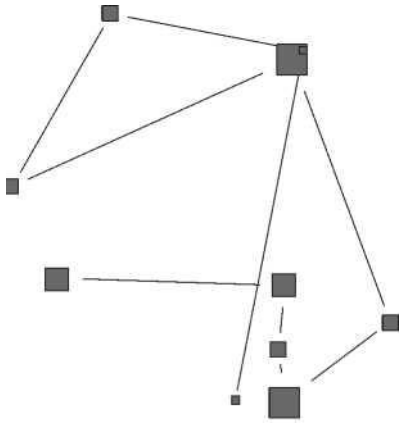


FIGURE 4.39 grid.edit() Output.
FIGURA 4.41 Ieșire grid.grill() .

FIGURA 4.40 Ieșire grid.rect() .

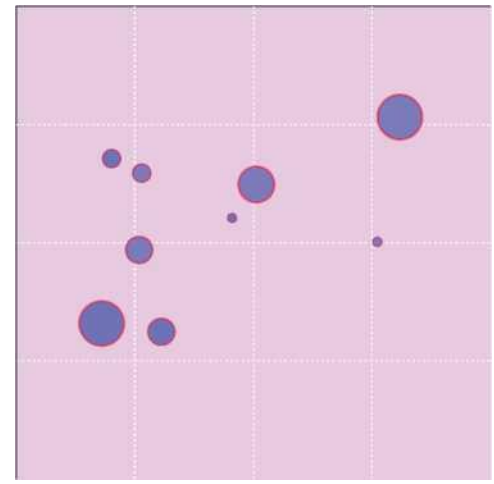


FIGURE 4.42 grid.draw() Output.

[4] CRAN package {lattice}. Available at http://cran.cnr.berkeley.edu

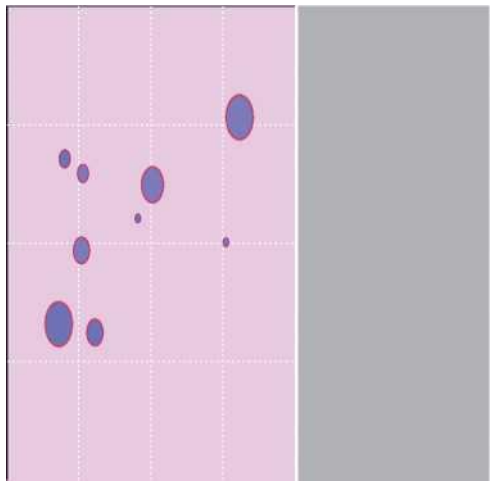


FIGURA 4.43 Ieșire dev.new() .

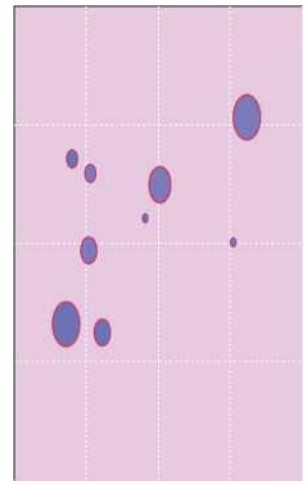


FIGURA 4.44 Ieșire grid.newpage() .

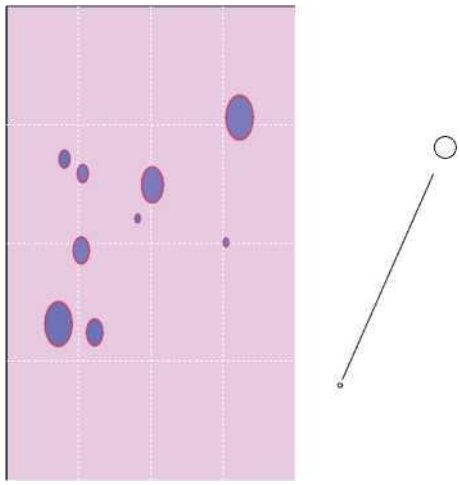


FIGURA 4.45 Ieșire grid.draw() .

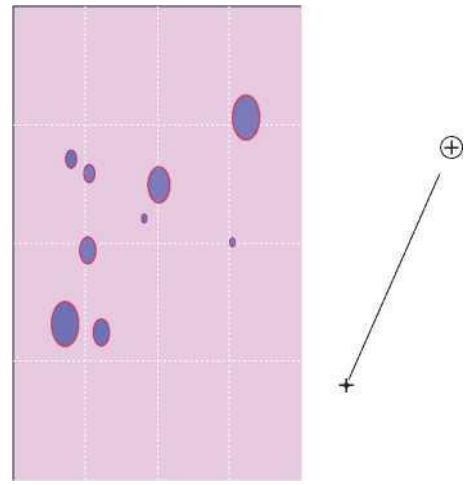


FIGURA 4.46 grid.points() Ieșire.

Murrell (2006) a recomandat următoarele funcții utile (Tabelul 4.2) care oferă - capacitatea de a interacționa cu grobs din grila de formă.*().

Aplicații la investigații biostatistice și epidemiologice

unele aplicații ale graficii grid pentru investigațiile biostatistice și epidemiologice din viața reală pentru a prezenta tehnicile de grafică grilă.

FUMI DE EŞAPAMENT DE MOTOR DIN ARDEREA ETANOLULUI

Această subsecțiune ilustrează o aplicație a graficului grilei – diagramă Trellis din pachetul latice (în CRAN). [4]

TABELUL 4.2 Câteva funcții utile pentru lucrul cu Grobs

FUNCȚIE DE FUNCȚIONARE CU IEȘIRE	DESCRIERE	FUNCȚIE DE LUCRARE CU GROBS
grid.get()	Returnează o copie a unuia sau mai multor grob-uri	getGrob()
grid.edit()	Modifică unul sau mai multe grobs	editGrob()
grid.add()	Adaugă un grob la unul sau mai multe grob	addGrob()
grid.remove()	Îndepărtează unul sau mai multe grobs	removeGrod()
grid.set()	Înlocuiește unul sau mai multe grobs	setGrob()

Descrierea anchetei:

Combustibilul etanol a fost ars într-un motor cu un singur cilindru. Pentru diferite setări ale raportului de compresie și echivalență a motorului, au fost înregistrate emisiile de oxizi de azot. A fost construit un cadru de date cu 88 de observații pe următoarele trei variabile:

NOx Concentrația oxizilor de azot (NO și NO_2) în micrograme/J
C Raportul de compresie al motorului
E Raportul de echivalență (o măsură a bogăției amestecului de combustibil de aer și etanol)

Autori:

Documentare: Wright, K.
Sursa: Brinkman (1981) .
Referință: Cleveland, William S. (1993). *Vizualizarea datelor.* Summit, NJ: Hobart Press.

```
>
>  # H_etanol
>  install.packages(„zăbrele")
>  bibliotecă ("zăbrele")
>
>  # Cele 88 de seturi de date din cadrul de date etanol pot fi inspectate de
>  # scoaterea datelor folosind: > etanol
>
>  ## Construirea funcțiilor panoului din mers
>  EE <- equal.count(etanol$E, număr=9, suprapunere=1/4)
> xyplot(NOx ~ C | EE, date = etanol,
+ prepanel = function(x, y) prepanel.loess(x, y, span = 1),
+ xlab = "Raport de compresie", ylab = "NOx (micrograme/J)",
+ panou = function(x, y) {
+ panel.grid(h=-1, v= 2)
+ panel.xyplot(x, y)
+ panel.loess(x,y, span=1)
+ },
+ aspect = "xy")
>  # Ieșire: Figura 4.47 etanol-1
```

[4] CRAN package {lattice}. Available at http://cran.cnr.berkeley.edu

```
>
>
>   # Se potrivește la suprafață loess cu cadru de sârmă (vezi Figura 4.48 ).
>   cere(statistici)
>   cu (etanol, {
+ eth.lo <- loess(NOx ~ C * E, span = 1/3, parametric = "C",
+ drop.square = "C", family="simetric")
+ et.marginal <- list(C = seq(min(C), max(C), lungime.out = 25),
+ E = seq(min(E), max(E), lungime.out = 25))
+ eth.grid <- expand.grid(eth.marginal)
+ eth.fit <- prezice(eth.lo, eth.grid)
+ wireframe(eth.fit ~ eth.grid$C * eth.grid$E,
+ nuanță=ADEVĂRAT,
+ ecran = listă (z = 40, x = -60, y=0),
+ distanta = .1,
+ xlab = "C", ylab = "E", zlab = "NOx")
+ })
>   # Ieșire: Figura 4.48 etanol -2
>
```

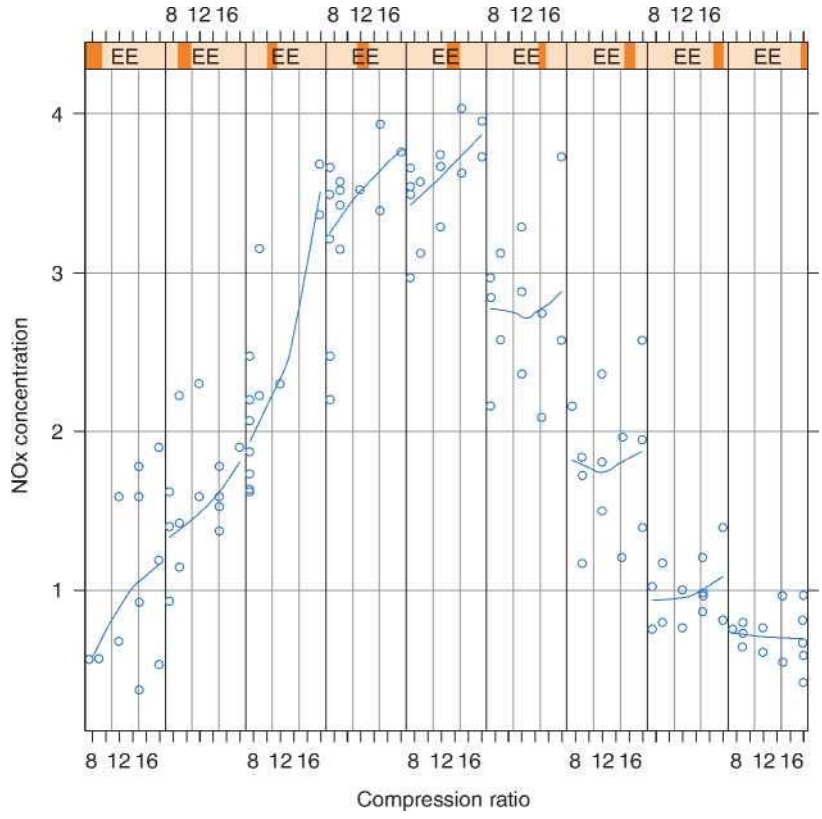


FIGURE 4.47 ethanol-1: Grid lattice panel.xyplot() of NOx concentration vs. compression ratio C, for varying fuel mixture equivalence ratios E.

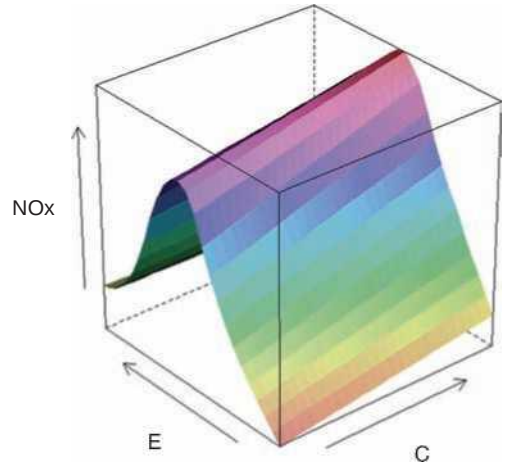


FIGURE 4.48 ethanol-2: Wireframe loess surface fit of NOx concentration vs. compression ratio C, for varying fuel mixture equivalence ratios E.

EPIDEMIOLOGIA DE MEDIU: CONDIȚII ATMOSFERICE ÎN NEW YORK CITY

Descrierea anchetei:

Măsurătorile zilnice ale concentrației de ozon, vitezei vântului, temperaturii și radiației solare în New York City din mai până în septembrie 1973.

Un cadru de date cu 111 observații asupra următoarelor patru variabile:

ozon — Concentrația medie de ozon (măsurători pe oră) în părți pe miliard *de radiații* — Radiația solară (de la 08:00 la 12:00) în *temperatura langleys* —Temperatura maximă zilnică în grade Fahrenheit
vântul — Viteza medie a vântului (la 07:00 și la 10:00) în mile pe oră

Autori:

Documentare: Wright, K.
Sursa: Bruntz, SM, Cleveland, WS, Kleiner, B. și Warner, JL (1974). Dependența - ozonului ambiental de radiația solară, vânt, temperatură și înălțimea de amestecare. În *Simpozionul despre difuzia atmosferică și poluarea aerului* (pp. 125–128). Boston, MA: Societatea Americană de Meteorologie.
Referință: Cleveland, WS (1993). *Vizualizarea datelor* . Summit, NJ: Hobart Press.

Acest exemplu de epidemiologie de mediu arată trei moduri grafice diferite de prezentare a patru factori de mediu (ozon, radiații, temperatură și condiții de vânt) utilizând grafica grilă – diagrama Trellis din pachetul de zăbrele (în CRAN). [4]

> **Exemplu:** {latice} de mediu (Murrell, 2006)

```
> install.packages(„zăbrele")
> bibliotecă ("zăbrele")
>
> # Cele 111 seturi de date din mediul cadru de date pot fi inspectate de
> # scoaterea datelor folosind: > mediu
```

```
>
>   # splom() grafic al mediului de cadru de date
>   # Scatter plot matrice cu linii loess
> splom(~de mediu,
+ panel=funcție(x,y){
+       panel.xyplot(x,y)
+       panel.loess(x,y)
+}
+ )
>   # Funcția splom(), în rețeaua pachetului , atrage dispersia condiționată
>   # grafice matrice și diagrame de coordonate paralele [5]
>   # ieșire: Figura 4.49 de mediu -1
>   # Slot condiționat similar cu Figura 5.3 din Cleveland
```

[5] CRAN site for splom() {lattice}:
http://127.0.0.1:27736/library/lattice/html/splom.html.

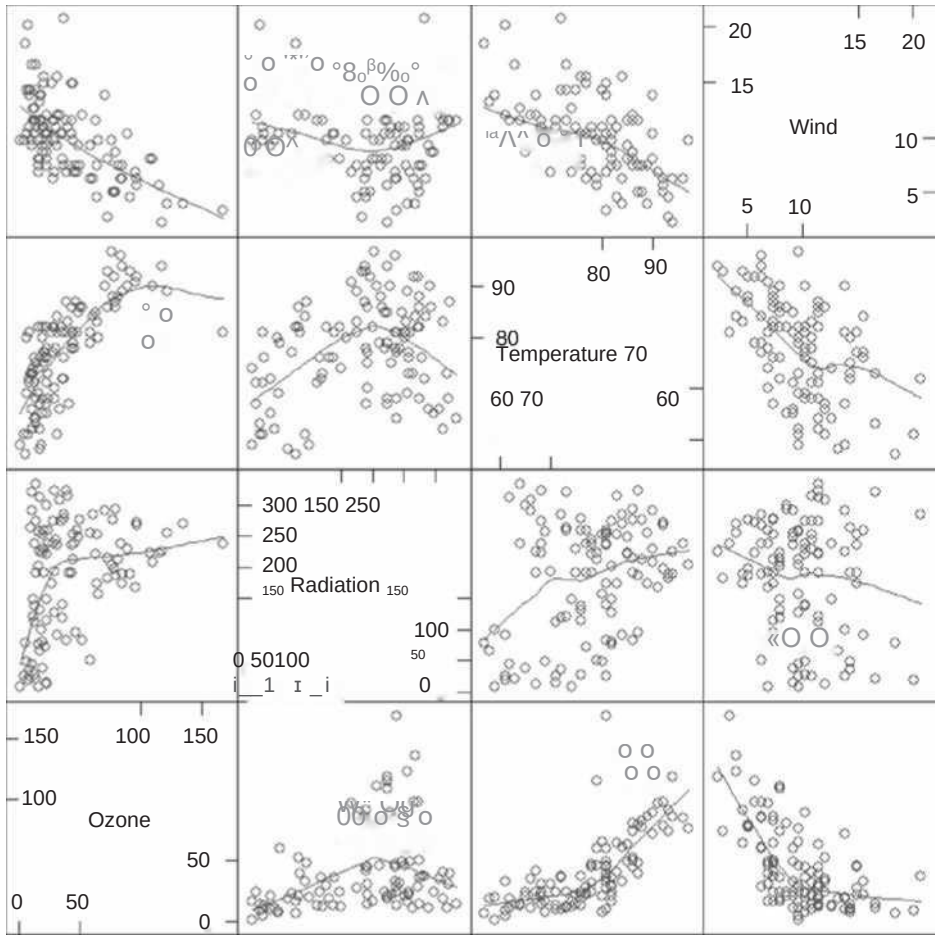


FIGURE 4.49 environmental-1: Grid lattice splom() plot of ozone, radiation, temperature, and wind conditions.

[5] CRAN site for splom() {lattice}: http://127.0.0.1:27736/library/lattice/html/splom.html.

```
>   atașează (de mediu)
>   Temperatura <- equal.count(temperature, 4, 1/2)
>   Vânt <- equal.count(vânt, 4, 1/2)
> xyplot((ozon^(1/3)) ~ radiație | Temperatura * Vânt,
+ aspect=1,
+ prepanel = funcția (x, y)
+ prepanel.loess(x, y, span = 1),
+ panou = funcția (x, y){
+ panel.grid(h = 2, v = 2)
+ panel.xyplot(x, y, cex = .5)
+ panel.loess(x, y, span = 1)
+ },
+ xlab = "Radiația solară (langleys)",
+ ylab = "Ozon (rădăcină cubă ppb)")
>   # ieșire: Figura 4.50 de mediu- 2
>   detașează()
>
```

> **Exemplu:** mediu **-2** {zăbrele}

```
>   install.packages(„zăbrele")
>   bibliotecă ("zăbrele")
>
>   # Slot condiționat similar cu Figura 5.3 din Cleveland
>   atașează (de mediu)
>   Temperatura <- equal.count(temperature, 4, 1/2)
>   Vânt <- equal.count(vânt, 4, 1/2)
> xyplot((ozon^(1/3)) ~ radiație | Temperatura * Vânt,
+ aspect=1,
+ prepanel = funcția (x, y)
+ prepanel.loess(x, y, span = 1),
+ panou = funcția (x, y){
+ panel.grid(h = 2, v = 2)
+ panel.xyplot(x, y, cex = .5)
+ panel.loess(x, y, span = 1)
+ },
+ xlab = "Radiația solară (langleys)",
+ ylab = "Ozon (rădăcină cubă ppb)")
>   # ieșire: Figura 4.50 de mediu - 2
>   detașează()
>
```

> **Exemplu:** mediu **-3** {zăbrele}

```
>   install.packages(„zăbrele")
>   bibliotecă ("zăbrele")
>
>   # Afișare similară folosind funcția coplot
```

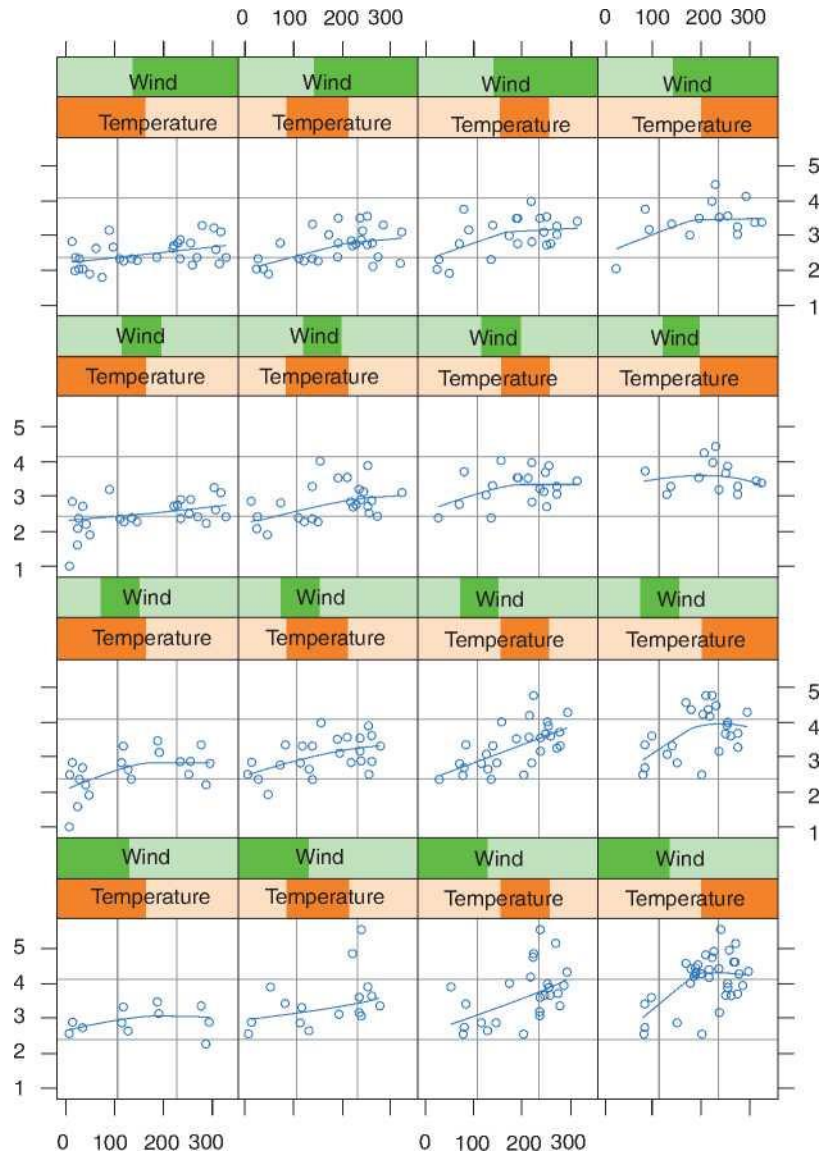


FIGURE 4.50 environmental-2: Grid lattice panel.xyplot() of ozone, radiation, temperature, and wind conditions.

```
> cu(mediul,{
+ coplot((ozon^.33) ~ radiație | temperatură * vânt,
+ număr=c(4,4),
+ panel = function(x, y, ...) panel.smooth(x, y, span = .8, ...),
+ xlab="Radiția solară (langleys)",
+ ylab="Ozon (rădăcină cubă ppb)")
+ })
>   # Ieșire: Figura 4.51 de mediu-- 3
>
```

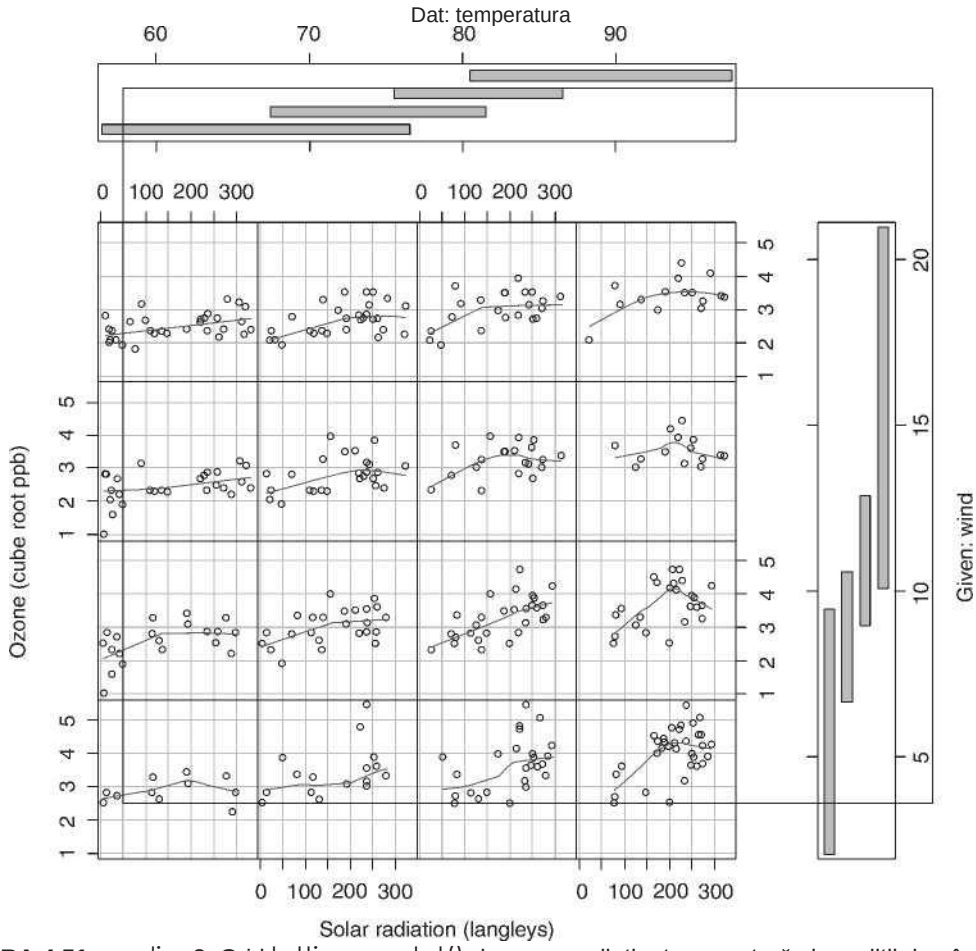


FIGURA 4.51 mediu -3: Grid lattice coplot() de ozon, radiație, temperatură și condiții de vânt.

EPIDEMIOLOGIA MEDIULUI II

Această subsecțiune folosește „ Lattice-p.44: B_06_levelplot—Level plots and contour plots" pentru a produce graficele dorite.

Descrierea anchetei:

Aceasta este aceeași investigație ca și exemplele din subsecțiunea precedentă.

Abordare analitică:

Pentru a afișa variațiile din mediul setului de date multivariat prin desenarea graficelor de nivel de culoare și a graficelor de contur, utilizând următoarea funcție din pachetul latice : contourplot(x, data, ...)

```
>
>  cere(statistici)
>  atașează (de mediu)
>  ozo.m <- loess((ozon^(1/3)) ~ vant * temperatura *
```

```
+ radiație, parametrică = c(„radiație", „vânt"),
+                                interval = 1, grad = 2)
>   w.marginal <- seq(min(vânt), max(vânt), lungime.out = 50)
>   t.marginal <- seq(min(temperatura), max(temperatura),
+                          lungime.out = 50)
>   r.marginal <- seq(min(radiatie), max(radiatii),lungime.out= 4)
>   wtr.marginal <- list(vent = w.marginal, temperatura =
+                          t.marginal, radiație = r.marginal)
>   grid <- expand.grid(wtr.marginal)
>   grid[, "potrivire"] <- c(predict(ozo.m, grid))
> contourplot(potrivire ~ vânt * temperatură | radiație, date = grilă,
+                          tăieturi = 10, regiune = TRUE,
+                  xlab = „Viteza vântului (mph)",
+                  ylab = "Temperatura (F)",
+                  main = "Cube Root Ozon (rădăcină cubă ppb)")
>   # Ieșire Figura 4.52 de mediu -4
```

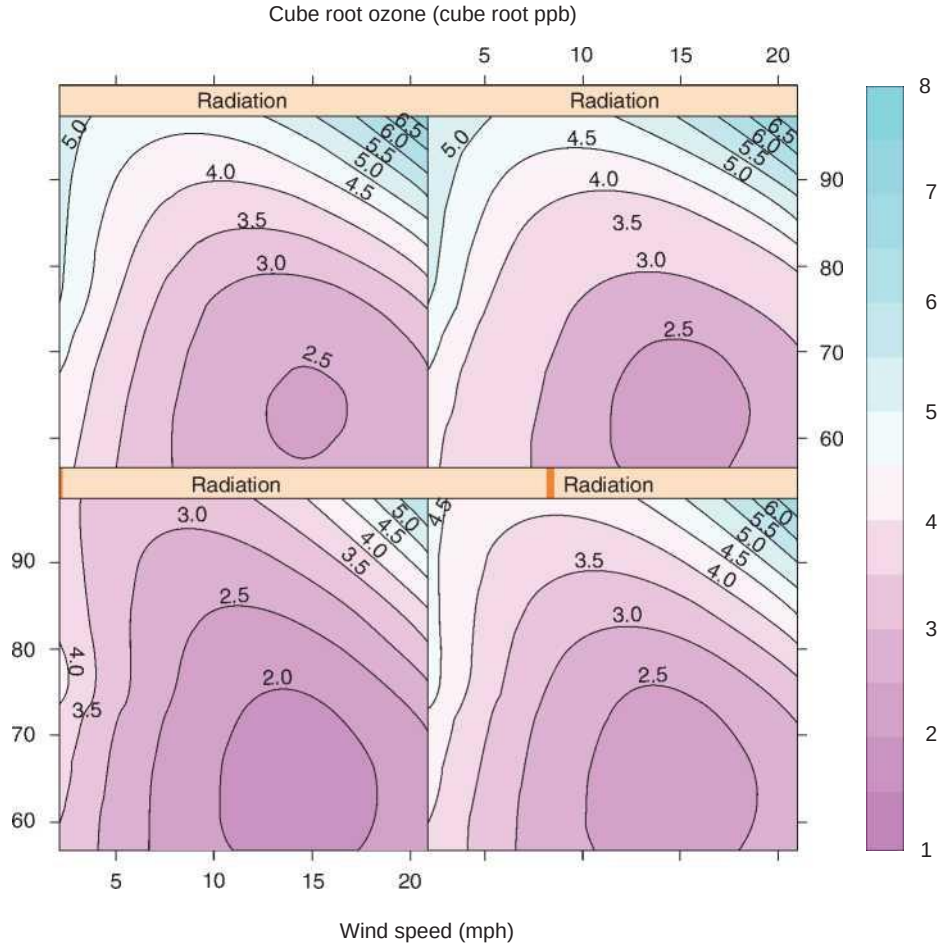


FIGURE 4.52 environmental-4: Grid lattice contourplot() of ozone, radiation, temperature, and wind conditions.

Această subsecțiune folosește grafica grilă – diagramă în spalier din pachetul lattice (în CRAN) [4] pentru a produce o diagramă în serie de timp (Sarkar, 2011a) .

Descrierea anchetei:
Datele din Connecticut Tumor Registry prezintă numere ajustate în funcție de vârstă ale incidenței cancerului de piele melanom la 100.000 de persoane în statul american Connecticut pentru anii 1936-1972. Constă dintr-un cadru de date cu 37 de observații pe următoarele două variabile:

anul - anii 1936 până în 1972
incidență — Rata cancerului de melanom la 100.000 de locuitori

Autor(i):
Documentare: Wright, K.
Sursa: Houghton, A., Munster, EW, & Viola, MV (1978). Creșterea incidenței melanomului malign după vârfurile activității petelor solare. *Lancet, 8* , 759–760.
Referință: Cleveland, WS (1993). *Vizualizarea datelor* . Summit, NJ: Hobart Press.

```
>   # Diagrama seriei temporale (Figura 3.64 din Cleveland).
> xyplot(incidență ~ an,
+ date = melanom,
+ aspect = "xy",
+ panou = funcția (x, y)
+ panel.xyplot(x, y, type="o", pch = 16),
+ ylim = c(0, 6),
+ xlab = "An",
+ ylab = "Incidență"
+)
> # Ieșire: Figura 4.53 melanom
```

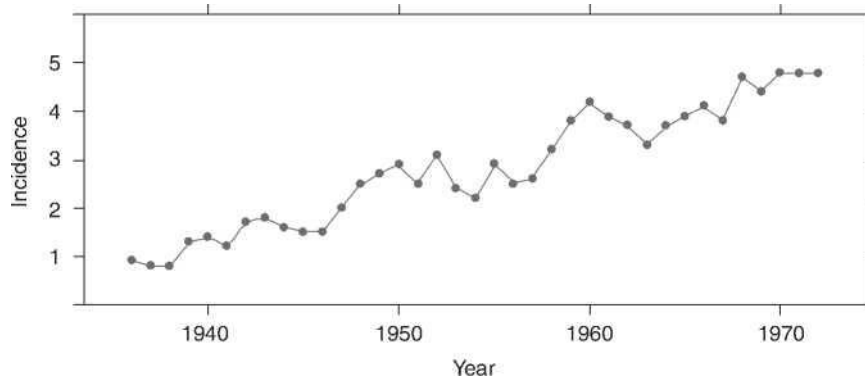


FIGURA 4.53 melanom grid lattice panel.xyplot(x, y) time-series plot.

EXEMPLE DIN PACHETUL CRAN {latticeExtra}

În 2011, Sarkar (2011b) a publicat pachetul CRAN latticeExtra , care conține o serie de funcții de grafică grilă care pot suporta afișări grafice pentru seturi de date biostatistice raportate în activități de cercetare și investigație în epidemiologie și sănătate publică. Câteva exemple de

aplicații ale acestor funcții la seturile de date raportate sunt prezentate aici.

■ **Exemplul 4.12:** Grafică grilă în R: gvhd10 în latticeExtra

Set de date: gvhd10 - Date de citometrie în flux (FCM) din cinci probe de la un pacient

Notă: FCM este o tehnică de examinare și *numărare* a particulelor microscopice, cum ar fi *cromozomii* și celulele, prin suspendarea acestora într-un curent de fluid și trecerea lor printr-un aparat electronic de detectare. Permite analiza multiparametrică simultană a caracteristicilor fizice și chimice de până la mii de particule pe secundă. FCM este folosit în diagnosticarea tulburărilor de sănătate, în special a cancerelor de sânge, dar are multe alte aplicații atât în cercetare, cât și în practica clinică.

Descrierea anchetei:
Datele FCM sunt înregistrate din probe de sânge prelevate de la un pacient cu leucemie înainte și după transplantul alogen de măduvă osoasă (o procedură de transplant în care pacientul primește celule stem de la un donator compatibil genetic, dar nu identic). Datele acoperă cinci vizite.

Utilizare: date(gvhd10)

Formatul datelor: un cadru de date cu 113.896 de observații pe următoarele opt variabile:

FSC.H	valorile înălțimii de împrăștiere
SSC.H	valorile înălțimii dispersării laterale
FL1.H	intensitatea (înălțimea) în canalul FL1
FL2.H	intensitatea (înălțimea) în canalul FL2
FL3.H	intensitatea (înălțimea) în canalul FL3
FL2.A	intensitatea (aria) în canalul FL2
FL4.H	intensitatea (înălțimea) în canalul FL4
Zile	un factor cu niveluri -6 0 6 13 20 27

Referință: Brinkman, RR, și colab. (2007) . Citometrie în flux cu conținut ridicat și analiză temporală a datelor pentru definirea unei semnături celulare a bolii grefă-versus-gazdă. *Biologia transplantului de sânge și măduvă , 13* (6), 691–700.

Segmentul de cod R pentru analiză este următorul:

```
> install.packages(„laticeExtra")
> bibliotecă ("laticeExtra")
Se încarcă pachetul necesar: RColorBrewer
Se încarcă pachetul necesar: zăbrele
> date (gvhd10)
>
> # Cele multe mii de seturi de date din cadrul de date gvhd10 pot fi
> # inspectat prin scoaterea datelor folosind: > gvhd10
```

```
>

> histogramă(~log2(FSC.H) | Zile, gvhd10, xlab = "înregistrare înainte
+ Scatter", tip = "densitate", nint = 50, aspect = c(2, 4))
>   # Ieșire: Figura 4.54 .
```

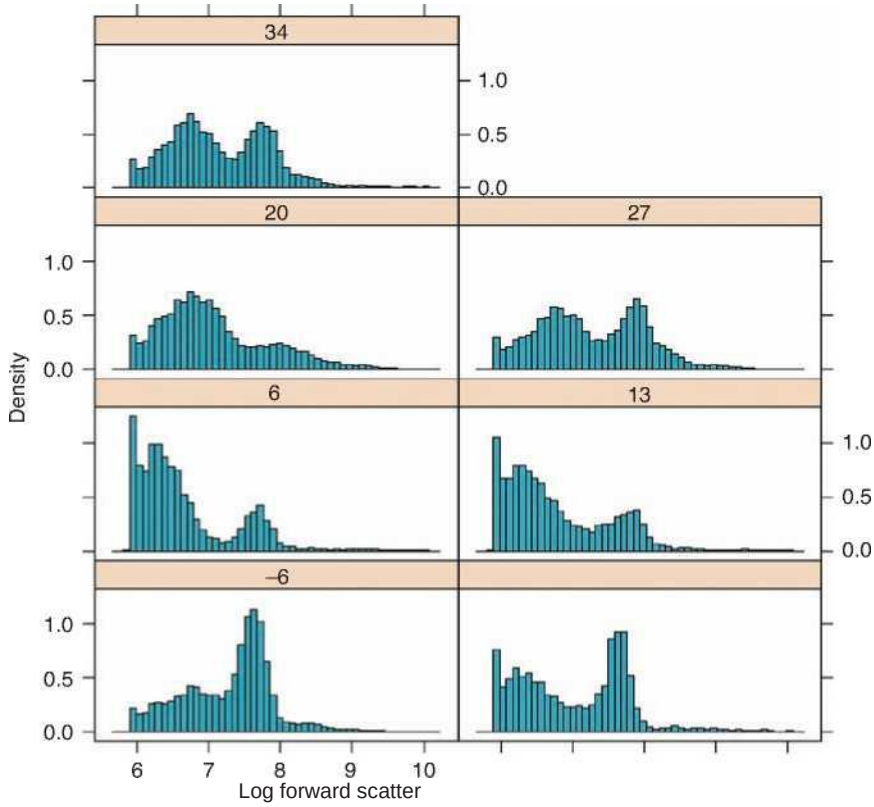


FIGURA 4.54 Grafica grilă R : Histogramă pentru ghvd10 în latticeExtra cu factorul Zile etichetat deasupra fiecărui panou.

pixmapGrob(x): UN PLOT GRIL SPECIAL PENTRU TOȚI UTILIZATORII R

Pentru a afișa însemnele R , funcția grilă pixmapGrob() poate fi utilizată pentru a crea un grob dintr-un obiect pixmap .

Descrierea funcției grid: Folosind funcția grid R pixmapGrob() , în pachetul RGraphics , se poate crea un grob dintr-un obiect pixmap și produce însemnele familiare ale programului R însuși.

Utilizare:

pixmapGrob(pic, x = 0,5, y = 0,5, scară = 1,

raster = FALSE, unghi = 0, vp = NULL, ...)

Argumente:

Pic	obiect Pixmap
X	*x* -unitate
Y	*y* -unitate

Scala Scala

Raster Raster

Unghi Unghi

Vp Viewport

... Parametri grob opționali, transmiși la imageGrob() sau rasterGrob()

Detalii:

Funcție foarte primitivă, folosind imageGrob sau rasterGrob de la R Graphics (R > = 2,11)

Valoare:

Un gTree de clasă „pixmap", cu lățimea și înălțimea naturală în puncte

R necesar este:

```
>  install.packages(„RGgraphics")
>  bibliotecă (pixmap)
>  bibliotecă (RGrafice)
>  x <- read.pnm(system.file("pictures/logo.ppm",
+ pachet="pixmap")[1])
>  g1 <- pixmapGrob(x)
>  dev.new(width=g1$width/72, height=g1$height/72)
>  grid.draw(g1)
>  # Ieşire: Figura 4.55.
```



FIGURE 4.55 Drawing R using the grid function pixmapGrob(x) in the package RGraphics.

Întrebări de revizuire pentru Secțiunea 4.2

1. (a) Ce este sistemul de grafică grilă?
   (b) Denumiți câteva pachete din acest sistem.
2. (a) Contrastați funcțiile de reprezentare din grafica grilă {latice} cu graficele de bază (tradiționale).
   (b) Dați un exemplu de dotplot() .
3. (a) Ce două tipuri de aranjamente sunt disponibile pentru mai multe diagrame pe o singură pagină folosind diagrame cu zăbrele?
   (b) Dați un exemplu de fiecare tip.
4. La adnotarea unui diagramă în zăbrele din mai multe panouri, pentru care funcție este disponibilă
   (a) controlând scalarea și dimensiunea panourilor?
   (b) controlul conținutului în fâșiile unei parcele cu zăbrele ?
   Dați un exemplu pentru fiecare.
5. (a) În grafica grilă, ce este un grob?
   (b) Scrieți un segment de cod R care definește un grob.
6. În grafica grilă, pentru a desena cercuri în locații și raze date, care funcție este disponibilă
   (a) pentru a crea un obiect?
   (b) pentru a produce o ieșire a obiectului creat?
   Dați un exemplu pentru fiecare.
7. În grafica grilă, descrieți funcțiile:
   (a) viewport()
   (b) pushViewport()
   Dați un exemplu pentru fiecare.
8. În lucrul cu grobs în grafica grilă, descrieți funcțiile:
   (a) getGrob()
   (b) grid.get()
   Dați un exemplu pentru fiecare.
9. (a) Descrieți pachetul CRAN latticeExtra .
   (b) Comparați cele două funcții map() și mapplot() din acest pachet.
10. (a) Descrieți pachetul CRAN RGraphics .
    (b) Utilizați funcția grid pixmapGrob() din acest pachet pentru a crea un obiect pixmap la alegere.

Exerciții pentru Secțiunea 4.2

1. rețelei pachetului de grafică grilă pentru a produce grafică Trellis. În mediul R , executați următorul segment de cod, comandă prin comandă. Pentru *fiecare* comandă:
   (a) Explicați acțiunea așteptată de la comandă.
   (b) După executarea comenzii, descrieți rezultatele și ieșirile.
2. Comparați afișarea finală a ieșirii obținute în Exercițiul 1 cu tipul de diagramă cu zăbrele: numele funcției R pentru producerea fiecărui tip de diagramă este afișat

pe banda de deasupra fiecărui diagramă. Pentru a obține aceeași ieșire, rescrieți segmentul de cod pentru a obține un program mai succint care va îndeplini aceeași sarcină.

```
>   bibliotecă (zăbrele)
>   trellis.device()
>   bibliotecă (grilă)
> x <- 1:5
> x
>   y <- 1:5
>   y
>   g <- factor (1:5)
>   g
>   tipuri <- c("barchart", "bwplot", "densityplot", "dotplot",
+           „histogramă", „qqmath", „stripplot", „qq",
+           „xyplot", „levelplot", „contourplot",
+           „cloud", „wireframe", „splom", „paralel")
>   tipuri
>   unghi <- seq(0, 2*pi, lungime=21)[-21]
>   unghi
>   xx <- cos(unghi)
>   xx
>   yy <- sin(unghi)
>   da
>   gg <- factor(rep(1:2, fiecare=10))
>   gg
>   aaa <- seq(0, pi, lungime=10)
>   aaa
>   xxx <- rep(aaa, 10)
>   xxx
>   aaa <- rep(aaa, fiecare=10)
>   aaa
>   zzz <- sin(xxx) + sin(aaa)
>   zzz
>   doplot <- function(name, ...) {
+ do.call(nume,
+    list(..., scale=list(draw=FALSE), xlab=NULL, ylab=NULL,
+           bandă=funcție(which.panel, ...) {
+           grid.rect(gp=gpar(fill="grey90")); grid.text(nume)
+                                                  }))}
>   plot <- vector("listă", 15)
>   complot
>   plot[[1]] <- doplot(„barchart", y ~ g | 1)
>   complot[[1]]
>   plot[[2]] <- doplot("bwplot", yy ~ gg | 1,
+                   par.settings=list(box.umbrella=list(lwd=0.5)))
>   complot[[2]]
>   plot[[3]] <- doplot("densityplot", ~ yy | 1)
```

```
>   complot[[3]]
>   plot[[4]] <- doplot(„dotplot", y ~ g | 1)
>   complot[[4]]
>   plot[[5]] <- doplot(„histogramă", ~ yy | 1)
>   complot[[5]]
>   plot[[6]] <- doplot("qqmath", ~ yy | 1)
>   complot[[6]]
>   plot[[7]] <- doplot(„stripplot", yy ~ gg | 1)
>   complot[[7]]
>   plot[[8]] <- doplot("qq", gg ~ yy | 1)
>   complot[[8]]
>   plot[[9]] <- doplot("xyplot", xx ~ yy | 1)
>   complot[[9]]
>   plot[[10]] <- doplot("levelplot", zzz ~ xxx + yyy | 1,
+ colorkey=FALSE)
>   complot[[10]]
>   plot[[11]] <- doplot("contourplot", zzz ~ xxx + yyy | 1,
+                          etichete=FALSE, tăieturi=8)
>   complot[[11]]
>   plot[[12]] <- doplot(„nor", zzz ~ xxx + yyy | 1, zlab=NULL,
+              zoom=0,9, par.settings=list(box.3d=list(lwd=0,01)))
>   complot[[12]]
>   plot[[13]] <- doplot("wireframe", zzz ~ xxx + yyy | 1,
>                  zlab=NULL, zoom=0,9, drape=TRUE,
+ par.settings=list(box.3d=list(lwd=0.01)), colorkey=FALSE)
>   complot[[13]]
>   complot[[13]]
>   plot[[14]] <- doplot("splom", ~ data.frame(x=xx[1:10],
+                          y=aa[1:10]) | 1, pscale=0)
>   complot[[14]]
>   plot[[15]] <- doplot(„paralel", ~ data.frame(x=xx[1:10],
+                          y=aa[1:10]) | 1)
>   complot[[15]]
>
>   grid.newpage()
>   pushViewport(viewport(layout=grid.layout(4, 4)))
>   pentru (i în 1:15) {
+ pushViewport(viewport(layout.pos.col=((i - 1) %% 4) + 1,
+                       layout.pos.row=((i - 1) %/% 4) + 1))
+ print(plot[[i]], newpage=FALSE,
+                       panel.width=list(1.025, "inci"),
+                       panel.height=list(1,025, „inci"))
+ popViewport()}
>
>   popViewport()
```

REFERINȚE

Altman, DG (1991). Exercițiul 11.2. În *Statistică practică pentru cercetarea medicală* . Boca Raton, FL: Chapman & Hall.

Baptiste, A. (2012, 14 februarie). Pachetul CRAN „gridExtra”. Preluat de pe http://cran.cnr .berkeley.edu/

Brinkman, ND (1981). Combustibil cu etanol - Un studiu al motorului cu un singur cilindru al eficienței și al emisiilor de evacuare. *SAE Transactions, 90* , 1410–1424.

Dalgaard, P. (2002). *Statistici introductive cu R* (Seria Statistică și Calcul). New York, NY: Springer.

Mittal, HV (2011). *R graphs carte de bucate* . Birmingham, Marea Britanie: PACKT Publishing.

Murrell, P. (2006). *Grafică R* (Seria informatică și analiză a datelor). Boca Raton, FL: Chapman & Hall/CRC. Preluat de la http://www.stat.auckland.ac.nz/~paul/grid/grid. html

Olsen, AH, Njor, SH, Vejborg, I., Schwartx, W., Dalgaard, P., Jensen, MB, . . . Lynge, E. (2005).

Mortalitatea prin cancer de sân la Copenhaga după introducerea screening-ului mamografiei. *British Medical Journal* , *330* , 220–222.

Sarkar, D. (2011a). Lattice Graphics—Pachetul „latice”, Versiunea 0.20-0. Preluat de la http://r-forge.r-project.org/projects/lattice/

Sarkar, D. (2011b). Pachetul CRAN „laticeExtra”. Preluat de pe http://cran.cnr.berkeley.edu/; http://latticeextra.r-forge.r-project.org/

Venables, WN, Smith, DM și echipa principală de dezvoltare R. (2005). *O introducere în R* (ed. Rev.). Bristol, Marea Britanie: Network Theory Limited.

FIVE

Probabilitate și statistică în biostatistică

INTRODUCERE

De ce să studiezi probabilitatea? După cum sa discutat în capitolul 2 , s-au stabilit diferite școli de inferență biostatistică. Aceste școli, sau paradigme, nu se exclud reciproc; metodele care funcționează satisfăcător sub o singură paradigmă oferă adesea interpretări atractive și în alte paradigme. Cele două paradigme principale utilizate sunt biostatistica bayesiană și biostatistica frecventistă.

Fundamentul pentru inferența biostatistică, sub orice abordare, este **teoria probabilității** . Deși această teorie, adesea considerată o ramură a matematicii, nu este punctul central al acestei cărți, este benefic să examinăm conceptele sale fundamentale, deoarece acestea se aplică analizei biostatistice în cadrul disciplinelor epidemiologie și sănătate publică.

Conceptul de probabilitate este familiar investigatorilor epidemiologici și lucrătorilor din sănătatea publică și medicina preventivă. De exemplu, cineva poate afla de la un lucrător din domeniul sănătății că cineva are „90% șanse" de a contracta malarie în anumite condiții de mediu sau poate auzi un oncolog spunând că un pacient are „50-50 șanse" de a supraviețui unui anumit diagnostic de cancer. .

După cum arată aceste exemple informale, probabilitățile sunt adesea exprimate în termeni de procente sau fracții. Probabilitatea de apariție a unui eveniment este un număr cuprins între 0 (pentru nicio șansă) și 1 (pentru o certitudine absolută). Cu cât este mai probabil ca evenimentul să se producă, cu atât probabilitatea se apropie de 1; cu cât este mai puțin probabil ca evenimentul să aibă loc, cu atât probabilitatea este mai apropiată de 0.

În epidemiologie și sănătate publică/medicină preventivă, investigatorii trebuie adesea să se întrebe dacă rezultatele observate inițial ale muncii lor ar fi putut apărea din pură întâmplare sau dacă alți factori definitivi au fost operați pentru a produce eforturile observate. De exemplu, dacă 5 din 10 pacienți sunt vindecați de o anumită boală după ce primesc un anumit tratament, întrebarea poate fi pusă după cum urmează:

> S-ar fi produs probabil această rată de vindecare dacă pacienții nu ar fi primit tratamentul sau rezultatul poate fi considerat o dovadă a unui efect de vindecare adevărat al tratamentului?

Aplicații ale teoriei probabilității (Dalgaard, 2002; Daniel, 2005; Kolmogorov, 1964 ; Triola și Triola, 2006) sunt utile în abordarea unor astfel de întrebări.

5.1 TEORIILE PROBABILITĂȚII

Ce este probabilitatea?

Probabilitatea a fost luată în considerare în cel puțin două moduri: subiectiv și obiectiv.

Probabilitatea subiectivă consideră probabilitatea ca o măsură a încrederii pe care o are cineva în adevărul unei anumite propoziții. Nu depinde de niciun proces sau de repetabilitate. Acest concept permite determinarea probabilității unui singur eveniment care se poate întâmpla o singură dată: de exemplu, probabilitatea ca un tratament complet pentru cancer să fie găsit în 25 de ani.

Biostatisticienii subscriu în general la celălalt concept de probabilitate, **probabilitatea obiectivă** , care este el însuși împărțit în două categorii: frecvență clasică și frecvență relativă.

Probabilitatea clasică a fost dezvoltată în secolul al XVII-lea pentru a rezolva probleme în jocurile de noroc, cum ar fi jocurile de cărți sau lansarea unui zar cu șase fețe:

- Dacă o carte este extrasă la întâmplare dintr-un pachet de 52 de cărți de joc (bine amestecate), probabilitatea de a trage asul de pică este 1/52, probabilitatea de a trage un diamant este 13/52 și așa mai departe.
- Dacă se aruncă un zar corect, atunci probabilitatea de a obține fața cu șase puncte este 1/6, probabilitatea de a obține fața cu patru puncte este 1/6 și așa mai departe.

Folosind acest concept, se iau în considerare doar evenimentele *la fel de probabile* – iar prezența fizică a unui pachet de cărți sau a unui zar nu este necesară.

Următoarea este o bună definiție a probabilității clasice:

> Când un eveniment poate avea loc în N moduri la fel de probabile (și se exclud reciproc) și dacă m dintre aceste evenimente au loc întotdeauna cu o proprietate E , atunci probabilitatea de apariție a lui E este m / N .

Notația comună pentru această definiție este

$$P (E) = m / N \tag{5,1-1}$$

care se citește ca „probabilitatea lui E este m împărțită la N ”.

Abordarea **probabilității de frecvență relativă** depinde de repetabilitatea unui proces, precum și de capacitatea de a enumera repetițiile și de numărul de ori în care are loc un eveniment de interes. Astfel, pentru a afirma probabilitatea de a observa apariția unei caracteristici E a unui eveniment, se folosește următoarea definiție a probabilității de frecvență relativă:

> Dacă un proces se repetă *de n* (un număr mare de) ori și dacă un eveniment cu caracteristica E are loc *de m* ori, atunci frecvența relativă a apariției lui E , fiind m / n , va fi *aproximativ* egală cu probabilitatea E .

Această definiție poate fi exprimată ca

$$P (E) = m / n \tag{5,1-2}$$

cu avertismentul că m / n este doar o *estimare* a *lui* $P (E)$.

Proprietățile de bază ale probabilității

Întregul sistem al teoriei probabilităților se bazează pe următoarele trei proprietăți:

Proprietatea 1 : Într-un experiment sau proces cu *n rezultate sau evenimente* **care se exclud reciproc** : $E_1, E_2, E_3, \ldots, E_n$, probabilitatea unui anumit eveniment E_i primește un număr nenegativ:

$$P(E_i) \geq 0 \qquad (5,1\text{-}3)$$

Astfel, toate evenimentele trebuie să aibă o probabilitate nenegativă. Conform acestei definiții, două evenimente se exclud reciproc dacă și numai dacă nu pot avea loc simultan.

Proprietatea 2 : Suma probabilităților rezultatelor este egală cu 1:

$$P(E_1) + P(E_2) + P(E_3) + \ldots + P(E_i) + \ldots + P(E_n) = 1 \quad (5.1\text{-}4)$$

Aceasta se numește **proprietatea exhaustivității.** Este necesar ca observatorul unui - proces probabil să permită toate evenimentele posibile, în timp ce proprietatea care se exclude reciproc garantează că cele *n* evenimente nu se suprapun (adică, două dintre ele nu pot avea loc simultan).

Proprietatea 3 : Pentru oricare două evenimente care se exclud reciproc E_j și E_k, probabilitatea apariției oricărui eveniment este egală cu suma probabilităților lor individuale:

$$P(E_j \text{ sau } E_k) = P(E_j) + P(E_k) \qquad (5.1\text{-}5)$$

- **Exemplul 5.1:** Calcularea probabilității de a avea fete și băieți într-o familie

Să presupunem că, în cadrul unei familii, a avea fete și a avea băieți este la fel de probabil și că genul oricărui copil nu este influențat de genul celorlalți copii din familie. Dacă un cuplu căsătorit, John și Mary Smith, plănuiește să aibă patru copii, aflați probabilitatea ca aceștia să aibă

- 4 fete (si fara baieti)
- 3 fete si 1 baiat
- 2 fete si 2 baieti
- 1 fata si 3 baieti
- 4 băieți (și fără fete)

Soluţie:

Pentru a rezolva această problemă, folosiți ecuația (5.1-1) : $P(E) = m / N$.

Pentru a găsi *m* și *N* pentru fiecare dintre cele cinci cazuri, ar trebui mai întâi să enumerați toate scenariile la fel de probabile pentru cazul de a avea patru copii. Se poate începe prin a examina **spațiul eșantion** - toate combinațiile posibile ale modurilor în care pot apărea patru copii - care este așa cum se arată în Tabelul 5.1 .

Acest spațiu eșantion arată că există 16 rezultate posibile la fel de probabile pentru domnul și doamna Smith. Ele reprezintă cele 16 rezultate *diferite posibile : N* = 16.

(i) Caracteristica E_1 = 4 fete (și fără băieți):

Doar 1 rezultat posibil corespunde exact la 4 fete și niciun băiat:

Cazul #1, deci *m* = 1.

Prin urmare, prin ecuația (5.1-1) : $P(E1) = m / N = 1/16$, *sau 0,0625,* sau 6,25%.

TABELUL 5.1 Spațiu eșantion: Toate combinațiile posibile de a avea patru copii

G = FATA, B = BĂIAT

Caz #:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
primul copil	G	G	G	G	B	G	G	G	B	B	B	G	B	B	B	B
al 2-lea copil	G	G	G	B	G	G	B	B	B	G	G	B	G	B	B	B
al 3-lea copil	G	G	B	G	G	B	G	B	G	B	G	B	B	G	B	B
al 4-lea copil	G	B	G	G	G	B	B	G	G	G	B	B	B	B	G	B

(ii) Caracteristica $E_2 = 3$ fete și 1 băiat:
Doar 4 rezultate posibile corespund exact la 3 fete și 1 băiat:
Cazurile #2, #3, #4 și #5, deci $m = 4$.
Prin urmare, prin ecuația (5.1-1) : $P(E2) = m / N = 4/16$, sau 1/4, sau 0,25, sau 25%.

(iii) Caracteristica $E_3 = 2$ fete și 2 băieți:
Doar 6 rezultate posibile corespund exact la 2 fete și 2 băieți:
Cazurile #6, #7, #8, #9, #10 și #11, deci $m = 6$.
Prin urmare, prin ecuația (5.1-1) : $P(E3) = m / N = 6/16$, sau 3/8, sau $_{0,375}$, sau 37,5%.

(iv) Caracteristica $E_4 = 1$ fată și 3 băieți:
Doar 4 rezultate posibile corespund exact la 1 fată și 3 băieți:
Cazurile #12, #13, #14 și #15, deci $m = 4$.
Prin urmare, prin ecuația (5.1-1) : $P(E4) = m / N = 4/16$, sau 1/4, sau $_{0,25}$, sau 25%.

(v) Caracteristica $E_5 = 4$ băieți (și fără fete):
Doar un singur rezultat posibil corespunde exact la 4 băieți și nicio fete:
Cazul #16, deci $m = 1$.
Prin urmare, prin ecuația (5.1-1) : $P(E5) = m / N = 1/16$, sau 0,0625, $_{\text{sau}}$ 6,25%.

Observatii:
Observați cei doi pași în determinarea probabilității oricărei caracteristici:
Pasul 1: *Enumerați* spațiul eșantion al tuturor rezultatelor posibile.
Pasul 2: Pentru o caracteristică specifică, căutați în întreg spațiul eșantion și *enumerați* numărul de rezultate posibile care poartă caracteristica specifică.

Acum putem verifica rezultatele cu privire la cele trei proprietăți de bază ale probabilității:

Proprietatea I: Deoarece fiecare dintre probabilitățile $P(E_i)$ este pozitivă, în mod clar

$$P(E_i) \geq 0, \text{ după cum este necesar.}$$

Proprietatea 2: $P(E_1) + P(E_2) + P(E_3) + P(E_4) + P(E_5)$

$= 1/16 + 4/16 + 6/16 + 4/16 + 1/16$

$= (1 + 4 + 6 + 4 + 1)/16$

$= 16/16$

$= 1$ sau 100%, după caz

Proprietatea 3: Condiția $P(E_j \text{ sau } E_k) = P(E_j) + P(E_k)$ este în mod clar satisfăcută, după cum se poate observa prin examinarea tabelului 5.1 și a valorilor celor cinci probabilități calculate $P(E_i) | i = 1, 2, 3, 4, 5$.

Conceptul de probabilitate poate fi discutat în termenii următoarelor proprietăți speciale (fiecare definită într-o subsecțiune următoare):

- Probabilitatea pentru evenimente complementare
- Probabilitate condiționată
- Probabilitate comună
- Regula înmulțirii pentru probabilități
- Regula adunării pentru probabilități
- Independenta si dependenta de evenimente
- Probabilitate marginală

PROBABILITATEA DE EVENIMENTE COMPLEMENTARE

Dacă se dorește să găsească probabilitatea ca un anumit eveniment *E* să *nu* se producă, atunci se spune că se caută complementul **evenimentului** E, notat cu $\underline{E}$.

Deoarece trebuie să fie adevărat că fie evenimentul *E* are loc, fie evenimentul *E* nu are loc, se poate scrie:

$$P(E) + P(\underline{E}) = P(\text{toate aparițiile posibile}) = 1$$

astfel încât

$$P(\underline{E}) = 1\ P(E) \qquad (5,1\text{-}6)$$

care poate fi considerată definiția complementului **evenimentului** E.

- **Exemplul 5.2:** Utilizarea complementului pentru a calcula probabilitatea de a avea fete și băieți într-o familie

Folosind din nou cazul domnului și doamnei Smith, să presupunem că aceștia și-ar dori totuși să aibă patru copii, dar numai cu condiția:

- Fără fete (adică doar băieți)
- 2 fete si 2 baieti
- Toate cu excepția a 2 fete și 2 băieți
- Fără băieți (adică doar fete)

Care sunt probabilitățile pentru fiecare dintre aceste condiții?

Soluţie:

Aici putem folosi ecuația (5.1-6) : $P(\underline{E}) = 1\ P(E)$.

(i) P (un eveniment pentru „fără fete”)

$= P(E_g)$ where E_g = the “no girls” event

$=1\ P(E_g)$ by Equation (5.1-6), where E_g = the “any girls

$= 1\ \{P(E_1) + P(E_2) + P(E_3) +$ from Example 5.1

$P(E_4)\}$

$= 1\ \{1/16 + 4/16 + 6/16 + 4/16\}$ from Example 5.1(i), (ii), (iii), (iv)

$= 1\ 15/16$

$= 1/16$ sau 0,0625 sau 6,25%

(ii) P (un eveniment pentru „doar 2 fete și 2 băieți”)

= 6/16 sau 3/8 sau 0,375 sau 37,5% din exemplul A1(iii)

(iii) *P* un eveniment pentru „toți, cu excepția a 2 fete și a 2 băieți")

= 1 *P* (*E*) unde E = „doar 2 fete și 2 băieți"

= 1 – 6/16 din Exemplul A1(iii) și partea (ii) din acest exemplu

= 10/16

= 5/8 sau 0,625 sau 62,5%

(iv) *P* (un eveniment pentru „fără băieți") = *P* (E_1) unde E_1 = evenimentul „1 băiat" și E_1 = un eveniment *fără* băiat

= 1 – 15/16, prin numărarea evenimentelor din Tabelul 5.1

= 1/16 sau 0,0625 sau 6,25%

Observatii:

1. Ambele cazuri (i) și (iv) au aceeași probabilitate, 0,0625%. Acest lucru implică faptul că este foarte puțin probabil – aproximativ 6% șanse – să nu existe fete sau băieți deloc.
2. Cazurile (ii) și (iii) sunt rezultate interesante. S-ar putea fi înclinați să creadă că, pentru o familie cu 4 copii, probabilitatea pentru un rezultat de 2 fete și 2 băieți ar fi de 50% - dar nu este cazul pentru 62,5% dintre rezultatele posibile.
3. Rețineți că, dacă familia are doar 2 copii, probabilitatea pentru un rezultat cu 1 fată și 1 băiat ar fi într-adevăr de 50%. (Arătați că acesta este într-adevăr cazul!)

PROBABILITATE CONDIȚIONATĂ

Având în vedere oricare două evenimente *A* și *B* , dacă se presupune mai întâi că evenimentul *A* **a avut deja loc** , atunci probabilitatea ca evenimentul *B* să se producă este **probabilitatea condiționată** *P* (*B* | *A*), care se citește ca probabilitatea ca *B* să se producă dată *A* , sau ca probabilitatea ca evenimentul *B* să se producă după ce evenimentul *A* a avut deja loc.

4. **Exemplul 5.3: Utilizarea** conceptului de probabilitate condiționată pentru a calcula probabilitatea ca o fiică, într-o familie de patru copii, să aibă o soră sau un frate

Folosim din nou cazul familiei de patru copii a domnului și doamnei Smith. Dacă această familie are deja o fiică, care sunt probabilitățile ca această fiică să aibă o soră sau să aibă un frate?

(i) Probabilitatea ca o fiică să aibă o soră:

Dacă *A* este fiică în această familie, pentru ca *A* să aibă o soră, trebuie să existe cel puțin 2 fete în setul de 4 copii din familia Smith. Prin urmare, probabilitatea ca *A* să aibă o soră este probabilitatea condiționată de a avea cel puțin 2 fete în setul de 4 copii.

Din Tabelul 5.1 , constatăm că există 16 rezultate posibile (*N* = 16). Prin enumerare, vedem că există 11 rezultate posibile cu cel puțin 2 fete fiecare: acestea sunt cazurile #1 până la #11 (*G* = 11). Prin urmare, probabilitatea condiționată, scrisă ca *P* (*G* | *N*), este

$$P(G \mid N) = 11/16 \text{ sau } 0{,}6875 \text{ sau } 68{,}75\%$$

Astfel, *A* are șanse de aproape 70% să aibă o soră.

(ii) Probabilitatea ca o fiică să aibă un frate:

Dacă *A* este fiică în această familie, pentru ca *A* să aibă un frate, trebuie să existe cel

puțin 1 fată și 1 băiat în setul de 4 copii din familia Smith. Prin urmare, probabilitatea ca *A* să aibă un frate este probabilitatea condiționată de a avea cel puțin 1 fată și 1 băiat în setul de 4 copii.

Din Tabelul 5.1 , constatăm că există 16 rezultate posibile ($N = 16$). Prin enumerare, constatăm că există 14 rezultate posibile cu cel puțin 1 fată și 1 băiat: acestea sunt cazurile #2 până la #15 ($B = 14$). Prin urmare, probabilitatea condiționată, scrisă ca $P(B \mid N)$, este

$$P(B \mid N) = 14/16 \text{ sau } 7/8 \text{ sau } 0{,}875 \text{ sau } 87{,}5\%$$

Astfel, *A* are șanse de aproape 90% să aibă un frate.

PROBABILITATE COMUNĂ

Având în vedere două variabile aleatoare *A* și *B* , **distribuția comună** pentru *A* și *B* definește probabilitatea evenimentelor definite atât în termenii lui *A* cât și ai *B* . Pentru două variabile aleatoare, aceasta produce o **distribuție bivariată** ; pentru orice număr de variabile aleatoare, aceasta produce o **distribuție multivariată** .

■ **Exemplul 5.4:** Utilizarea conceptului de probabilitate comună pentru a calcula probabilitățile unei fiice primul născut, într-o familie de 4 copii, având 2 surori și 1 frate sau 1 soră și 2 frați

Din nou, folosim cazul familiei de patru copii a domnului și doamnei Smith. Dacă primul născut din această familie este o fată, G1, întrebarea este: Care sunt probabilitățile ca G1 să aibă (i) 2 surori și 1 frate? (ii) 1 soră și 2 frați?

În acest exemplu, se presupune că faptul că o familie are o fată sau un băiat este un eveniment independent din punctul de vedere al probabilității. (În unele familii, factorii genetici, precum și alți factori, pot denatura șansele în favoarea unui sex sau altul; fetele sau băieții pot doar „a alerga în familie". În scopul acestor exemple, ignorăm totuși. acești factori.)

Din Tabelul 5.1 , vedem că din cele 16 rezultate posibile, există 8 în care primul născut este o fată: Cazurile #1, #2, #3, #4, #6, #7, #8 și #12.

(La fel, există 8 în care primul născut este un băiat.) Prin urmare, $N_g = 8$.

Din aceste 8 cazuri, doar 3 au alte 2 fete și 1 băiat: Cazurile #2, #3 și #4. Prin urmare, $G_{21} = 3$ și

$$P(G21 / Ng) = G21 / Ng = 3/8$$

De asemenea, din aceste 8 cazuri, doar 3 au încă 1 fată și 2 băieți: Cazurile #6, #7 și #9. Prin urmare, $G_{12} = 3$ și

$$P(G12 / Ng) = G12 / Ng = 3/8$$

Probabilitatea necesară este probabilitatea comună a acestor două evenimente. Deoarece acestea sunt evenimente independente, probabilitatea lor comună este suma probabilităților lor individuale. Fie această probabilitate comună $P[(G_{21}, G_{12})/ N_g]$. Apoi

$$P[(G_{21}, G_{12})/ N_g] = P(G_{21} / N_g) + P(G_{12} / N_g) = 3/8 + 3/8 = 6/8, \text{ sau } 3/4, \text{ sau } 75\%$$

Prin urmare, o fată prima-născută are șanse de 75% să aibă fie 2 surori și 1 frate, fie 1 soră și 2 frați.

REGULA MULTIPLICĂRII PENTRU PROBABILITĂȚI

În notație simbolică, regula înmulțirii probabilităților poate fi scrisă astfel: Pentru oricare

două evenimente *A* și *B,*

$$P(A \cap B) = P(A)\,P(B \mid A), \text{ dacă } P(A) \neq 0 \qquad (5.1\text{-}7A)$$

sau

$$P(A \cap B) = P(B)\,P(A \mid B), \text{ dacă } P(B) \neq 0 \qquad (5.1\text{-}7B)$$

în care simbolul ∩ este citit fie ca „și" sau „intersecție" (așa cum este folosit în teoria mulțimilor). Enunțul $A \cap B$ indică apariția comună a condițiilor *A* și *B.*

■ **Exemplul 5.5:** Folosirea regulii înmulțirii pentru probabilități pentru a calcula probabilitatea de a avea un fiu primul născut, într-o familie de 4 copii, care va avea după el 2 frați și 1 soră

Folosind cazul familiei de patru copii a domnului și doamnei Smith, presupunem că primul născut din această familie este un fiu, B1. Întrebarea este: Care sunt probabilitățile ca B1 să aibă 2 frați și 1 soră?

Tabelul 5.1 arată că din cele 16 rezultate posibile, $N = 16$, există 8 în care primul născut este un băiat: Cazurile #5, #9, #10, #11, #13, #14, #15, și #16. Prin urmare, $N_b = 8$.

Astfel, probabilitatea de a avea un fiu ca prim-născut este

$$P(B) = N_b / N = 8/16$$

Acum, dintre aceste 8 cazuri, 3 au 2 băieți și o fată: cazurile #13, #14 și #15. Prin urmare, $N_{bbg} = 3$, iar probabilitatea de a fi primul născut cu 2 frați și 1 soră este

$$P(A \mid B) = N_{bbg} / N_b = 3/8$$

Folosind regula înmulțirii pentru probabilități,

$$P(A \cap B) = P(B)\,P(A \mid B), \text{ dacă } P(B) \neq 0 \qquad (5.1\text{-}7B)$$

$$= (8/16) \times (3/8)$$

$$= 3/16 \text{ sau } 0{,}1875 \text{ sau } 18{,}75\%$$

Prin urmare, familia Smith se poate aștepta la o șansă ceva mai mică de 20% de a avea un fiu primul născut să fie urmat de 2 băieți și 1 fată.

REGULA DE ADULTARE PENTRU PROBABILITATI

Regula de adunare pentru probabilități spune că, având în vedere două evenimente *A* și *B* , probabilitatea ca evenimentul *A* sau *B* , sau ambele, să se producă este egală cu probabilitatea ca evenimentul *A* să se producă, **plus** probabilitatea ca evenimentul *B* să se producă, **mai puțin** probabilitatea. că aceste două evenimente au loc împreună.

În formă simbolică, această regulă poate fi exprimată ca

$$P(A \cup B) = P(A) + P(B)\,P(A \cap B) \qquad (5.1\text{-}8)$$

în care simbolul ∪ este citit fie ca „sau" sau „unire" (așa cum este folosit în teoria mulțimilor).

■ **Exemplul 5.6:** Utilizarea regulii de adunare pentru probabilități pentru a calcula probabilitatea de a avea fie un fiu primul născut, într-o familie de 4 copii, care va avea după el cel puțin 1 frate; sau un fiu primul născut, într-o familie de 4 copii, care va avea după el și cel puțin 1 soră

Tabelul 5.1 arată că din cele 16 rezultate posibile, $N = 16$, există 8 în care primul născut este un băiat: Cazurile #5, #9, #10, #11, #13, #14, #15, și #16. Prin urmare, $N_b = 8$, iar probabilitatea de a avea un fiu ca prim-născut este

$$P(N_b) = N_b / N = 8/16$$

Acum, dintre aceste 8 cazuri, 7 au cel puțin 1 băiat de urmat: Cazurile #9, #10, #11, #13, #14, #15 și #16. Prin urmare, $N_{bb} = 7$, iar probabilitatea de a fi primul născut cu 1 frate mai mic este

$$P(N_{bb} / N_b) = N_{bb} / N_b = 7/8$$

Folosind regula înmulțirii pentru probabilități:

$$P(N_{bb} \cap N_b) = P(N_b)\, P(N_{bb} \mid N), \text{ dacă } P(N) \neq 0 \text{ din ecuația (5.1-7B)}$$

$$= (8/16) \times (7/8)$$

$$= 7/16 \text{ sau } 0{,}4375 \text{ sau } 43{,}75\%$$

Desemnați această probabilitate ca $P(A)$.

Din Tabelul 5.1 , din cele 16 rezultate posibile, $N = 16$, există 8 în care primul născut este un băiat: Cazurile #5, #9, #10, #11, #13, #14, #15, și #16. Prin urmare, $N_b = 8$, iar probabilitatea de a avea un fiu ca prim-născut este

$$P(N_b) = N_b / N = 8/16$$

Acum, dintre aceste 8 cazuri, 7 au o fată de urmat: cazurile #9, #10, #11, #13, #14, #15 și #16. Acestea sunt aceleași șapte cazuri ca cele din prima parte a acestui exemplu.

Prin urmare, $N_{bg} = 7$, iar probabilitatea de a fi un fiu primul născut cu o soră mai mică

$$\text{va fi } P(N_{bg} \setminus N_b) = N_{bg} / N_b = 7/8.$$

Folosind regula înmulțirii pentru probabilități,

$$P(N_{bg} \cap N_b) = P(N_b)\, P(N_{bg} N_b), \text{ dacă } P(N_b) \neq 0 \text{ din ecuația (5.1-7B)}$$

$$= (8/16) \times (7/8)$$

$$= 7/16 \text{ sau } 0{,}4375 \text{ sau } 43{,}75\%$$

Desemnați această probabilitate ca $P(B)$.

Probabilitatea ca aceste două evenimente să se producă simultan este acele 7 cazuri indicate mai devreme, dintr-un total de 16 cazuri posibile. Prin urmare,

$$P(A \cap B) = 7/16$$

Acum aplicăm **regula adunării pentru probabilități** . Pentru ca ambele evenimente să apară simultan , probabilitatea este dată de

$$P(A \cup B) = P(A) + P(B)\, P(A \cap B) \qquad (5{,}1\text{-}8)$$

$$= 7/16 + 7/16\ 7/16$$
$$= 7/16, \text{ or } 0.4375, \text{ or } 43.75\%$$

Prin urmare, dacă familia Smith are cei patru copii ai săi, se poate aștepta să aibă o șansă mai mare de 40% să aibă un băiat ca prim-născut, cu un frate sau o soră care să-l urmeze.

INDEPENDENȚA ȘI DEPENDENTĂ DE ACURENȚE

Având în vedere două evenimente A și B , dacă evenimentul B a avut loc, dar nu are niciun efect asupra probabilității lui A (adică, probabilitatea evenimentului A rămâne neschimbată indiferent dacă are loc sau nu evenimentul B), atunci

$$P (A \mid B) = P (A) \tag{5.1-9}$$

Într-un astfel de caz, se spune că A și B sunt evenimente **independente** (în sensul probabilității).

Astfel, pentru două evenimente independente, regula înmulțirii [Ecuația (5.1-7B)] poate fi scrisă ca

$$P (A \cap B) = P (B) P (A \mid B), \text{ dacă } P (B) \neq 0 \tag{5.1-7B}$$

$$= P (B) P (A) \tag{5,1-10}$$

conform ecuației (5.1-9) . Adică:

$$P(A \cap B) = P (B) P (A), \text{ dacă } P (A) \neq 0, \text{ dacă } P (B) \neq 0 \tag{5,1-11}$$

Când două evenimente, A și B, cu probabilități diferite de zero sunt independente, fiecare dintre următoarele afirmații este adevărată:

$$P (A \mid B) = P (A); P (B \mid A) = P (B); P (A \cap B) = P (A) P (B)$$

Observatii:

1. În teoria probabilității, termenii *independenți* și *exclusivi reciproc* nu **înseamnă** neapărat același lucru.
2. Dacă evenimentele A și B nu sunt independente, se spune că sunt *dependente* .

3. **Exemplul 5.7:** Utilizarea independenței și dependenței de apariții pentru a calcula - probabilitățile de a avea o fiică întâi născută, într-o familie de 4 copii, care apoi are și surori

Revenim la cazul familiei de patru copii a domnului și doamnei Smith. Dacă primul născut din această familie este o fiică, G1, întrebarea este: Care sunt probabilitățile ca G1 să nu aibă (i) nicio soră, (ii) doar 1 soră, (iii) doar 2 surori sau (iv)) 3 surori?

În acest exemplu, se presupune din nou că faptul că o familie are o fată sau un băiat este un eveniment independent din punctul de vedere al probabilității. (După cum s-a menționat mai devreme, factorii genetici care „curg în familie” pot denatura aceste șanse.)

(i) Tabelul 5.1 arată că din cele 16 rezultate posibile, $N = 16$, există 8 în care primul născut este o fată: Cazurile #1, #2, #3, #4, #6, #7, #8, și #12. Prin urmare, $N_g = 8$ și

$$P (A) = P (N_g) = N_g / N = 8/16$$

Din aceste 8 cazuri, doar 1 nu mai are fete: Cazul #12. Prin urmare, $N_{g=0} = 1$ și

$$P (B) = P (N_{g=0}) = N_{g=0} / N_g = 1/8$$

Deoarece aceste două evenimente sunt independente,

$$P(A \cap B) = P(B)\,P(A)$$
$$= (1/8) \times (8/16)$$
$$= 1/16 \text{ sau } 0{,}0625 \text{ sau } 6{,}25\%$$

(ii) Tabelul 5.1 arată că din cele 16 rezultate posibile, $N = 16$, există 8 în care primul născut este o fată: Cazurile #1, #2, #3, #4, #6, #7, #8, și #12. Prin urmare, $N_g = 8$ și

$$P(A) = P(N_g) = N_g / N = 8/16$$

Dintre aceste 8 cazuri, doar 3 au încă o fată: cazurile #6, #7 și #8.
Prin urmare, $N_{g=1} = 3$ și

$$P(B) = P(N_{g=1}) = N_{g=1} / N_g = 3/8$$

Deoarece aceste două evenimente sunt independente,

$$P(A \cap B) = P(B)\,P(A)$$
$$= (3/8) \times (8/16)$$
$$= 3/16 \text{ sau } 0{,}1875 \text{ sau } 18{,}75\%$$

(iii) Tabelul 5.1 arată că din cele 16 rezultate posibile, $N = 16$, există 8 în care primul născut este o fată: Cazurile #1, #2, #3, #4, #6, #7, #8, și #12. Prin urmare, $N_g = 8$ și

$$P(A) = P(N_g) = N_g / N = 8/16$$

Dintre aceste 8 cazuri, doar 3 au încă două fete: cazurile #2, #3 și #4. Prin urmare, $N_{g=2} = 3$ și

$$P(B) = P(N_{g=2}) = N_{g=2} / N_g = 3/8$$

Deoarece aceste două evenimente sunt independente,

$$P(A \cap B) = P(B)\,P(A)$$
$$= (3/8) \times (8/16)$$
$$= 3/16 \text{ sau } 0{,}1875 \text{ sau } 18{,}75\%$$

(iv) Tabelul 5.1 arată că din cele 16 rezultate posibile, $N = 16$, există 8 în care primul născut este o fată: Cazurile #1, #2, #3, #4, #6, #7, #8, și #12. Prin urmare, $N_g = 8$ și

$$P(A) = P(N_g) = N_g / N = 8/16$$

Din aceste 8 cazuri, doar 1 mai are trei fete: Cazul #1. Prin urmare, $N_{g=3} = 1$ și

$$P(B) = P(N_{g=3}) = N_{g=2} / N_g = 1/8$$

Deoarece aceste două evenimente sunt independente,

$$P(A \cap B) = P(B)\,P(A)$$
$$= (1/8) \times (8/16)$$

Astfel, în această familie de 4 copii, dacă primul născut este fată, șansa pe care o va

$$= 1/16, \text{ or } 0.0675, \text{ or } 6.25\%$$

avea

- nu mai multe surori este 6,25%
- doar o altă soră este 18,87%
- doar încă 2 surori este 18,75%
- Încă 3 surori este 6,25%

PROBABILITATE MARGINALA

Dacă o variabilă poate fi împărțită în *n* categorii $A_1, A_2, A_3, \ldots, A_i, \ldots, A_n$ și o altă variabilă care apare în comun poate fi împărțită în *m* categorii $B_1, B_2, B_3, \ldots, B_j, \ldots, B_m$, atunci **probabilitatea marginală** a *lui* A_i, $P(A_i)$, este egală cu suma probabilităților comune ale lui A_i cu toate categorii de *B* ; adica:

$$P(A_i) = \sum P(A_i \cap B_j) \text{ , pentru toate valorile lui } j \qquad (5.1\text{-}12)$$

Următorul exemplu (Tabelul 5.2) ilustrează în continuare proprietățile speciale ale - probabilității marginale.

■ **Exemplul 5.8:** Utilizarea Ecuația (5.1-12) și datele privind frecvența consumului ilegal de droguri de către bărbați adulți (M) și femele (F), calculează probabilitatea marginală pentru bărbați, P(M)

Soluţie:

Variabila GEN este separată în două categorii: *M* și *F* .

Frecvența variabilă de viață a consumului de droguri ilegale este împărțită în trei categorii: de 1–19 ori (*A*), de 20–99 de ori (*B*) și ≥ 100 de ori (*C*).

Categoria *M* apare împreună cu toate cele trei categorii ale frecvenței variabile a consumului ilegal de droguri. Astfel, cele trei probabilități comune care pot fi comparate sunt

$P(M \cap A) = 32/111$; $P(M \cap B) = 18/111$; $P(M \cap C) = 25/111$

Pentru a obține probabilitatea marginală pentru bărbați, $P(M)$, aplicați ecuația (5.1-12)

:

$$P(A_i) = \sum P(A_i \cap B_j) \text{ , pentru toate valorile lui } j \qquad (5.1\text{-}12)$$

TABELUL 5.2 Frecvența consumului ilegal de droguri în funcție de sex

FRECVENȚA DE VIAȚĂ A CONSUMULUI ILEGAL DE DROGURI	BĂRBAȚI (M)	FEMEI (F)	TOTAL
1–19 ori (*A*)	32	7	39
de 20–99 de ori (*B*)	18	20	38
≥ 100 de ori (*C*)	25	9	34
Total	75	36	111

Sursa: Daniel (2005).

sau

$$P(M) = P(M \cap A) + P(M \cap B) + P(M \cap C)$$
$$= 32/111 + 18/111 + 25/111$$
$$= (32 + 18 + 25)/111$$
$$= 75/111 \text{ sau } 0{,}6757 \text{ sau } 67{,}57\%$$

Observație: același rezultat poate fi obținut și folosind totalul marginal pentru bărbați (75) ca numărător și numărul total de subiecți (111) ca numitor.

Calcule de probabilitate folosind R

Un număr de funcții utile în R sunt disponibile pentru calculele de probabilitate. Acestea sunt ilustrate în Exemplul 5.9 .

■ **Exemplul 5.9:** Funcțiile speciale R factorial() , choose() , sample () și prod() în calculele de probabilitate

(a) Funcția factorial()

Un epidemiolog testează efectele a cinci noi medicamente pentru cancer care urmează să fie administrate *secvenţial* unui grup de subiecţi cazaţi. Câte seturi de experimente (secvențe de testare) sunt necesare pentru a testa toate permutările posibile?

Pentru fiecare subiect de caz, secvența testului este

Drug1, Drug2, Drug3, Drug4, Drug5.

În mod clar, există cinci opțiuni pentru Drug1, plecare

4 opțiuni pentru Drug2,
3 opțiuni pentru Drug3,
2 opțiuni pentru Drug4 și
1 opțiune pentru Drug5.

Astfel, numărul total de secvențe de testare este 5 × 4 × 3 × 2 × 1, sau factorial 5 (factorial 5), scris de obicei ca 5! Acum, 5! = 120.

Următorul segment de cod R arată că funcţiile din baza pachetului sunt disponibile pentru astfel de calcule:

```
> install.packages(„bază'')
> bibliotecă ("bază")
> factorial(5) # ieșire:
[1]  120
> lfactorial(5) # Acesta este jurnalul natural al lui 5!
[1] 4,787492
> lovg(factorial(5)) # Este la fel ca lfactorial(5).
[1] 4,787492
```

Prin urmare, investigatorul epidemiologie ar trebui să planifice pentru 5! = 120 de secvențe de teste.

(b) Funcția alege()

Revenind la planificarea familială a domnului și doamnei Smith, dacă preferă să aibă 2 fete și 2 băieți, în câte moduri (ordinele de naștere a fetelor și băieților) se poate întâmpla acest lucru?

Prin enumerare folosind tabelul 5.1 :

Spațiu eșantion: Toate combinațiile posibile de a avea 4 copii: G = fată, B = băiat

Caz #:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
primul copil	G	G	G	G	B	G	G	G	B	B	B	G	B	B	B	B
al 2-lea copil	G	G	G	B	G	G	B	B	B	G	G	B	G	B	B	B
al 3-lea copil	G	G	B	G	G	B	G	B	G	B	G	B	B	G	B	B
al 4-lea copil	G	B	G	G	G	B	B	G	G	G	B	B	B	B	G	B

Se poate vedea că numărul de cazuri cu 2 G și 2 B este 6: Cazurile #6, #7, #8, #9, #10 și #11.

Folosind combinatoria, răspunsul este

$$
\begin{aligned}
{}_{n}C_{r} &= n\,!/\,r\,!\,(n-r)!\\
&= [\,n\,(n-1)\cdots(n-r+1)(n-r)!]/[\,x\,(x-1)\cdots 1][(nr)!]\\
&= [\,n\,(n-1)\cdots(n-r+1)]/[\,x\,(x-1)\cdots 1]
\end{aligned}
$$

Aici, $n = 4$ și $x = 2$, deci ${}_{n}C_{r} = {}^{4}C_{2} = 4 \times 3/2 \times 1 = 12/2 = 6$, confirmând rezultatul prin numărare.

Din nou, următorul segment de cod R arată că funcțiile din baza pachetului sunt disponibile pentru aceste calcule:

```
> install.packages(„bază")
> bibliotecă ("bază")
> alege(4, 2) # ieșire:
[1] 6
> lchoose(4, 2) # Acesta este logul natural al lui 4C2
[1] 1,791759
> log(choose(4, 2)) # Acesta este același cu lchoose(4, 2)
[1] 1,791759
```

(c) Eșantionul de funcție()

Eșantionarea aleatorie, cu și fără înlocuire, este un pas crucial în procesul de investigare epidemiologică a caracteristicilor de sănătate ale unei populații. Eșantionarea trebuie făcută deoarece pur și simplu nu este practic să se testeze întreaga populație din cauza numărului mare de indivizi din acea populație.

Să presupunem că, într-un oraș cu 1 milion de locuitori, fiecare cetățean este desemnat cu un număr, de la 1 la 1.000.000. Pentru testarea biostatistică, un lucrător din domeniul sănătății ar dori să selecteze aleatoriu (fără înlocuire) 3 seturi reprezentative de subiecți de caz, fiecare cu 5 persoane. Cum se poate face acest lucru?

Soluţie:

Funcția sample() , din baza pachetului , poate fi utilizată. Utilizarea acestei funcții (a se vedea documentația CRAN) ia următoarea formă pentru eșantionare fără înlocuire:

eșantion (x, dimensiune, înlocuire = FALSE, prob = NULL)

unde:

x = un număr întreg pozitiv (să zicem, 1)

dimensiune = un întreg nenegativ care oferă numărul de elemente din întreaga populație din care să aleagă

înlocuire = FALSE înseamnă eșantionare fără înlocuire; aceasta este valoarea implicită replace = TRUE face eșantionarea cu înlocuire

prob = un vector de greutăți de probabilitate pentru obținerea elementelor vectorului eșantionat (acest lucru poate fi omis)

Următorul segment de cod R execută calculele necesare:

> install.packages(„bază")
> bibliotecă ("bază")
> *# Comportamentul implicit al sample() este eșantionarea fără înlocuire;*
> *# aceasta înseamnă că fiecare persoană eșantionată nu va fi selectată de mai multe ori.*
> *# Pentru a obține 3 seturi reprezentative a câte 5 subiecte de caz fiecare, următoarele*
> *# calculul va fi executat de 3 ori:*
> eșantion (1:1000000, 5) *# ieșirea primului set de 5 subiecte de caz:*
> **1] 503512 33035 363755 527424 904495**
> eșantion (1:1000000, 5) *# ieșirea celui de-al doilea set de 5 subiecte de caz:*

[1] 853246 286220 211121 393481 842452

> eșantion (1:1000000, 5) *# ieșirea celui de-al treilea set de 5 subiecte de caz:*

1. **] 644293 870071 163122 153612 348948**

Observatii:

2. Pentru eșantionare aleatorie *cu* înlocuire, comanda este
   > eșantion (x, dimensiune, înlocuire = TRUE).
3. Pentru populațiile mari, eșantionarea fără înlocuire este tolerată. Pentru populații relativ mici, se poate alege să eșantioneze *cu* înlocuire.
4. Funcția sample() este potrivită pentru eșantionarea aleatorie. Pentru alte moduri de eșantionare mai restrictive (cum ar fi *eșantionarea echilibrată în cluster, eșantionarea echilibrată cu stratificare, eșantionarea echilibrată în două etape, eșantionarea în mai multe etape, eșantionarea minimă suport, eșantionarea multinomială, eșantionarea pivotală, eșantionarea Poisson, eșantionarea sistematică aleatorie, eșantionarea sistematică* etc.) , consultati documentatia CRAN pentru prelevarea pachetului .

(d) Funcția prod()

Dintr-o populație de pacienți de 20 de persoane, lucrătorul sanitar pregătește grupuri de câte 4 pentru teste clinice suplimentare. Câte grupuri pot fi combinate, fără a se referi la ordinea testării în cadrul fiecărui grup?

$$^{20}C_4 = 20!/\{4!(20\ 4)!$$

Soluţie:

Aceasta este o problemă elementară în combinatorică: determinarea numărului de - combinații posibile (fără ordonare) de grupuri de 4 dintr-o populație totală de 20. Răspunsul este $^{20}C_4$, care este dat de

$$^{20}C_4 = 20!/\{4!(20\ 4)!$$

În loc să evaluăm această expresie prin metode lungi, se poate folosi funcția prod() din baza pachetului . Utilizarea acestei funcții (vezi documentația CRAN) ia următoarea formă:

> prod(..., na.rm = FALSE)

pentru care argumentele sunt următoarele:

= un vector numeric sau complex sau logic

na.rm = logic. Ar trebui eliminate valorile lipsă ?

în R , calculul $^{20C4\ poate}{}_{fi}$ realizat într-unul din următoarele moduri, utilizând segmentele de cod indicate .

1. Într-un singur pas:
   > prod(20:1)/{prod(4:1)*prod((20 - 4):1)} # *20!/{4!(20 4)!*
   [1] 4845
2. Mai întâi, calculați fiecare dintre cei trei factoriali; și apoi combinați rezultatele:
   > prod(20:1) # *20!*
   [1] 2.432902e+18
   > prod(4:1) # *4!*
   [1] 24
   > prod((20 - 4):1) # *(20 4)!*
   [2] 2.092279e+13
   > 2.432902e+18/(24*2.092279e+13) # *Combinarea celor trei factoriali*
   [3] 4845
   Prin urmare, dintr-o populație de 20 de persoane, grupuri combinate de 4 pot fi formate în 4.845 de moduri.

Observatii:

1. Rezultatul se aplică pentru combinații de grupuri de câte 4 subiecți de caz fiecare și este exprimat ca $^{20}C_4$.
2. Dacă se ia în considerare ordonarea membrilor fiecărui grup, atunci numărul posibil de grupuri de 4 membri este: $^{20}P_4 = 20!/(20 – 4)!$

Următorul segment de cod poate fi utilizat pentru a calcula valoarea lui $^{20}P_4$:

> prod(20:1)/prod((20 4):1) **#** *Ieșire:*
[1] 116280

Acest rezultat poate fi obținut și prin următoarele considerații combinatorii:

(i) Pentru a constitui un grup de 4, sunt 4 posturi de ocupat: _____.
(ii) Primul post poate fi ocupat de oricare dintre cei 20 de persoane inițiale.
(iii) A doua pozitie de oricare dintre restul de 19,
(iv) A treia poziție de oricare dintre restul de 18, și
(v) A patra și ultima poziție de oricare dintre restul de 17.
(vi) Prin urmare, numărul total de astfel de grupuri de permutare posibile este

care este $^{20}P_4$. $20 \times 19 \times 18 \times 17 = 116{,}280$

Aplicații ale teoriei probabilităților în științele sănătății

Capitolul 2 , despre cercetare și proiectare în epidemiologie și sănătate publică, a arătat că aplicațiile conceptelor și teoriilor probabilităților sunt utilizate pe scară largă, în special pentru luarea deciziilor privind criteriile de diagnostic în medicina clinică și pentru testele de screening de sănătate în medicina preventivă. Epidemiologii și clinicienii ar beneficia de îmbunătățirea capacității lor cantitative de a prezice în mod eficient absența sau prezența unei anumite boli prin rezultatele testelor (negative sau pozitive) și starea simptomelor critice (absența sau prezența). În plus, acești profesioniști din domeniul sănătății ar fi interesați de informații cu privire la probabilitatea rezultatelor negative sau pozitive ale testelor și probabilitatea absenței sau prezenței unor simptome speciale în cazul subiecților cu sau fără anumite boli.

din capitolul 2 despre biostatistica bayesiană în epidemiologie a indicat că rezultatele unor astfel de teste de screening nu sunt întotdeauna infailibile. O anumită procedură de testare poate produce fals pozitive sau fals negative, așa cum se arată în tabelul 5.3 .

Aplicând teoria probabilității, se poate răspunde la următoarele patru întrebări atunci când se încearcă să evalueze aplicabilitatea și utilitatea rezultatelor testelor și a stării diagnosticului pentru a evalua dacă un subiect de caz are o boală specifică:

1. Dacă un subiect de caz nu are boala, **care este probabilitatea** de a obține un rezultat negativ al testului (sau absența unui simptom)?
2. Dacă un subiect de caz are boala, **care este probabilitatea** de a obține un rezultat pozitiv al testului (sau prezența unui simptom)?
3. Dacă testul de screening arată un rezultat negativ, sau testul de diagnostic arată absența unui simptom, **care este probabilitatea** ca subiectul cazului să nu aibă boala?
4. Dacă testul de screening arată un rezultat pozitiv sau testul de diagnostic arată prezența unui simptom, **care este probabilitatea** ca subiectul cazului să aibă boala?

ABORDAREA UTILIZAREA TEORIEI PROBABILITĂȚII

Pentru un eșantion mare de *n* subiecți de caz, se obține rezultatul prezentat în Tabelul 5.4

.

Tabelul 5.4 arată starea acestor *n* subiecți de caz în raport cu o anumită boală rezultată în urma unui test de screening diagnostic pentru identificarea persoanelor care

TABEL 5.3 Un tabel de decizie 2×2 care arată cele patru rezultate posibile dintr-un proces standard de testare clinică dihotomică

Rezultatele care reprezintă o eroare sunt scrise cu caractere aldine cursive.

	STAREA POPULAȚIEI EȘANȚĂ	
DECIZIA DE TESTARE	PACIENTUL ESTE INFECTAT	PACIENTUL NU ESTE INFECTAT
Pozitiv	Adevărat pozitiv	***Fals Pozitiv***
Negativ	***Fals Negativ***	Adevărat negativ

TABELUL 5.4 Rezultatele testelor de screening la *n* subiecți clasificați încrucișat în funcție de starea bolii

	BOALA		
REZULTATELE TESTULUI	PREZENT (D)	ABSENT ($\underline{D}$)	TOTAL
Pozitiv (T)	*p*	*q*	*p* + *q*
Negativ ($\underline{T}$)	*r*	*s*	*r* + *s*
Sumă	*p* + *r*	*q* + *s*	*n*

au boala. Intrările în celulă reprezintă numărul de subiecte de caz aparținând categoriilor definite de titluri de rând și coloană. Astfel, *p* este numărul de subiecți de caz care au boala și al căror rezultat al testului a fost pozitiv, iar *q* este numărul care nu au boala și al căror test a fost, de asemenea, pozitiv. Din informațiile din acest tabel, se pot deriva diferite estimări de probabilitate.

De exemplu:

Pentru a răspunde la întrebarea (1): Rețineți că estimarea probabilității condiționate este

$$P\left(\underline{T}\,/\,\underline{D}\right) = s\,/(\,q + s\,) \tag{5.1-13}$$

Acest raport este o estimare a **specificității** testului de screening diagnostic. Mai precis, *specificul* unui test de screening sau al unui simptom de diagnostic este probabilitatea unui rezultat negativ al testului (sau absența unui simptom), având în vedere absența reală a bolii.

Pentru a răspunde la întrebarea (2): Probabilitatea unui test pozitiv pentru subiecții de caz care au într-adevăr boala este dată de

$$P\,(\mathrm{T}|\mathrm{D}) = p\,/(\,p + r\,) \tag{5,1-14}$$

Acest raport este o estimare a **sensibilității** testului de screening diagnostic. Mai precis, *sensibilitatea* unui test de screening sau a unui simptom de diagnostic este probabilitatea unui rezultat pozitiv al testului (sau prezența unui simptom), având în vedere prezența reală a bolii.

Pentru a răspunde la întrebarea (3): Calculați probabilitatea condiționată:

$$P\,(\,D|T\,) = p\,/(\,p + q\,) \tag{5,1-15}$$

Acest raport este o estimare a probabilității ca subiectul să aibă boala, având în vedere că rezultatul testului de screening a fost pozitiv sau simptomul necesar a fost prezent. Se numește **valoarea predictivă pozitivă** a testului de screening diagnostic sau a unui simptom . Valoarea *predictivă pozitivă* a unui test de screening sau a unui simptom este probabilitatea ca un subiect de caz să aibă boala, având în vedere că subiectul are un rezultat pozitiv al testului de screening sau are simptomul necesar.

Pentru a răspunde la întrebarea (4): Calculați probabilitatea condiționată:

$$P\left(\underline{D}\,|\,\underline{T}\right) = s\,/(\,r + s\,) \tag{5.1-16}$$

Acest raport este o estimare a **valorii predictive negative** a testului de screening diagnostic sau a unui simptom. Valoarea *predictivă negativă* a unui test sau a unui simptom este

probabilitatea ca un subiect de caz să nu aibă boala, având în vedere că subiectul are un rezultat negativ al testului de screening sau nu prezintă simptomul necesar.

O ABORDARE FOLOSIND TEOREMA LUI BAYES

Teorema lui Bayes, care poate fi utilizată pentru a determina sensibilitatea și specificitatea testelor de screening, poate fi utilizată și pentru a obține estimările valorii predictive (pozitive și negative). În notarea tabelului 5.4 , următoarea afirmație oferă valoarea predictivă pozitivă a unui test sau simptom de screening de diagnostic:

$$P(D \mid T) = \frac{P(T \mid D)P(D)}{P(T \mid D)P(D) + P(T \mid D)P(D)} \quad (5.1\text{-}17)$$

Baza acestei ecuații poate fi găsită în regula înmulțirii pentru probabilități:

$$P(A \cap B) = P(B)\,P(A \mid B), \text{ dacă } P(B) \neq 0 \quad (5.1\text{-}7B)$$

sau

$$P(A \mid B) = P(A \cap B)/P(B), \text{ dacă } P(B) \neq 0$$

Prin urmare, probabilitatea condiționată

$$P(T \mid D) = P(T \cap D)/P(D), \text{ dacă } P(D) \neq 0$$

poate fi exprimat ca

$$P(TID)\,P(D) = P(T \cap D) = P(D \cap T)$$

astfel încât numărătorul lui Ecuația (5.1-17) reprezintă $P(D \cap T)$. *Se* poate observa că numitorul reprezintă pur și simplu $P(T)$, deoarece $P(D) + P(\underline{D}) = 1$. Acesta din urmă poate fi stabilit mai formal după cum urmează.

Evenimentele reprezentate de $P(D \cap T)$ și $P(\underline{D} \cap T)$ se exclud reciproc; adică au intersecție zero. Folosind regula adunării, Ecuația (5.1-8) , se poate scrie:

$$P(T) = P(D \cap T) + P(D \cap T) \quad (5.1\text{-}18)$$

Acum, după regula înmulțirii:

$$P(D \cap T) = P(T \mid D)\,P(D) \quad (5.1\text{-}19A)$$

şi

$$P(D \cap T) = P(T \mid D)\,P(\mathrm{D}) \quad (5.1\text{-}19B)$$

În cele din urmă, înlocuind ecuațiile (5.1-19A) și (5.1-19B) în ecuația (5.1-18) , obținem:

$$P(T) = \mathrm{P(T|D)\,P(D) + P(T|}\,\underline{\mathrm{D}}\,\mathrm{)\,P(}\,\underline{\mathrm{D}}\,\mathrm{)} \quad (5{,}1\text{-}20)$$

care este numitorul părții drepte a ecuației (5.1-13).

Observatii:

1. Numătorul ecuației (5.1-13) este egal cu sensibilitatea × rata (prevalența) bolii.
2. Numitorul ecuației (5.1-13) este egal cu sensibilitatea × rata bolii + (1 – sensibilitatea) × (1 – rata bolii).
3. Aceste relații permit ca valoarea predictivă pozitivă să fie calculată din sensibilitatea, specificitatea și rata bolii.

4. Un alt mod de a exprima teorema lui Bayes, analog cu ecuația (5.1-17) , este

$$P(\underline{D}\,|\,\underline{T}) = \{P(\underline{T}\,|\,\underline{D})\,P(\underline{D})\}/\{P(\underline{T}\,|\,\underline{D})\,P(\underline{D}) + P(\underline{T}\,|\,D)\,P(D)\} \qquad (5.1\text{-}21)$$

care permite să se calculeze o estimare a probabilității ca un subiect care este *negativ* la testul de screening (sau nu are simptom) să *nu* aibă boala – care este valoarea predictivă negativă a unui test de screening diagnostic sau a unui simptom.

În Exemplul ilustrativ 5.10 (Daniel, 2005) , teorema lui Bayes este utilizată pentru a calcula o valoare predictivă pozitivă a unui test de screening diagnostic.

■ **Exemplul 5.10:** Screening pentru boala Alzheimer

O echipă de investigatori epidemiologici clinici intenționează să evalueze o procedură de screening de diagnostic propusă pentru boala Alzheimer (AD). Procedura de testare a fost întreprinsă cu

(a) un eșantion aleatoriu de 900 de subiecți de caz care au avut AD și
(b) un eșantion independent aleatoriu de 1000 de subiecți de caz fără simptome ale bolii.

Aceste două probe au fost prelevate din populații care aveau cel puțin 65 de ani. Rezultatele sunt rezumate în Tabelul 5.5 .

Soluţie:
Estimarea sensibilității testului de screening poate fi obținută utilizând tabelul 5.4 și ecuația (5.1-14) :

$$P(\mathrm{T}|\mathrm{D}) = p/(p + r) \qquad (5,1\text{-}14)$$

$$= 872/(872 + 28)$$

$$= 872/900$$

$$= 0{,}9689$$

TABELUL 5.5 Rezumatul rezultatelor pentru testul de screening diagnostic al bolii Alzheimer

	DIAGNOSTICUL BOALA ALZHEIMER		
REZULTATELE TESTULUI	DA (D)	NU (D)	TOTAL
Pozitiv (T)	872	10	882
Negativ ($\underline{T}$)	28	990	1.018
Total	900	1.000	1.900

Specificitatea testului poate fi calculată utilizând tabelul 5.4 și ecuația (5.1-13) :

$$P(\underline{T}/\underline{D}) = s/(q+s) \qquad (5.1\text{-}13)$$

$$= 990/(10 + 990)$$

$$= 990/1000$$

$$= 0{,}9900$$

Pentru a calcula valoarea predictivă pozitivă a testului (adică, pentru a estima probabilitatea ca un subiect de caz care este pozitiv la testul de screening să aibă AD), se desprinde din Tabelul 5.5 :

$$P(T \mid D) = 872/900 = 0{,}9689$$

$$P(T \mid D) = 10/1000 = 0{,}0100$$

Înlocuind aceste valori în ecuația (5.1-17) , obținem:

$$P_1(D|T) \quad \overline{P(T \mid D)P(D) + P(T \mid D)P(D)} \qquad (5.1\text{-}17)$$

$$(0{,}9689)\,P(D)$$
$$(0{,}9689)\,P(D) + (0{,}0100)\,P(D)$$

Acum, $P(D)$, rata AD în populația generală relevantă, a fost estimată la 11,3% (Daniel, 2005) sau

$$P(D) = 0{,}1130$$

Dacă se acceptă această valoare pentru $P(D)$,
atunci

$$P(D) = 1\,P(D) = 1 - 0{,}1130 = 0{,}8870$$

Prin urmare,

$$P_1(D \mid T) = \frac{(0{,}9689)\,P(D)}{(0{,}9689)\,P(D) + (0{,}0100)\,P(D)} \qquad (5.1\text{-}17)$$

$$= \frac{(0.9689)(0.1130)}{(0.9689)(0.1130) + (0.0100)(0.8870)}$$

$$= 0{,}1095/0{,}1184$$
$$= 0{,}9248 \text{ sau } 92{,}48\%$$

Observatii:

1. Valoarea predictivă pozitivă a testului de screening este mai mare de 92%, o valoare foarte mare, ceea ce indică faptul că testul este fiabil.
2. Ecuația (5.1-22) arată că valoarea predictivă pozitivă a testului depinde de rata bolii în populația investigată: subiecți caz care au 65 de ani sau mai mult. Deoarece cele două eșantioane independente ale testului au fost prelevate de la două populații diferite, trebuie să se obțină o estimare independentă a $P(D)$.

Statistici rezumative tipice în biostatistică: intervale de încredere, teste de semnificație și bunătatea potrivirii

În biostatistică, **statisticile rezumative** sunt adesea folosite pentru a rezuma un set de observații , astfel încât să exprime cea mai mare cantitate cât mai succint posibil. Biostatisticienii pot descrie observațiile ca:

- O măsură a locației sau *a tendinței centrale* , cum ar fi mediana, media sau modul
- O măsură a dispersiei biostatistice, cum ar fi nivelul de încredere, varianța, abaterea standard (SD) sau intervalul
- O măsură a formei distribuției, cum ar fi normalitatea, asimetria sau curtoza
- O măsură a dependenței biostatistice, cum ar fi un coeficient de corelație, coeficienți de regresie sau altele asemenea, dacă se măsoară mai mult de o variabilă
- vizuale oferă adesea o imagine de ansamblu biostatistică și vizuală a unui eșantion, folosind o histogramă și/sau grafice cu puncte, casete, medie, percentile și SD pentru a prezenta datele într-o manieră grafică semnificativă vizual.

O colecție comună de statistici de ordine utilizată ca statistici rezumative este **rezumatul cu cinci numere** , care este o biostatistică descriptivă care oferă informații despre un set de observații. Este format din cele mai importante cinci percentile ale eșantionului:

1. Eșantionul minim
2. Quartila inferioară sau **primul quartila**
3. Mediana
4. Quartila superioară sau **a treia quartila**
5. Eșantionul maxim

Următorul este un exemplu de statistici rezumate folosind R .

> **Exemplul 5.11:** Statistici rezumate folosind R

Calculați rezumatul cu cinci numere în limbajul de programare R folosind funcția fivenum() . Când este aplicată unui vector, funcția summary() afișează rezumatul cu cinci numere *împreună* cu media (care nu face ea însăși parte din rezumat):

```
> x <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10) # Primele 10 numere naturale
> fivenum(x) # Ieșire: rezumatul cu cinci numere
[1]  1,0 3,0 5,5 8,0 10,0
> summary(x) # Ieșire: rezumatul statisticilor
Min.   Primul Qu.    Media media    a  3-a Qu.    Max.
>  .00    3.25    5.50    5.50    7.75   10.00
>  # Încercați cu un alt set de numere:
>  x <- rnorm(1000) # Luând în considerare 1000 de numere extrase aleatoriu din
>                   # distribuția normală standard
>  cincinum(x)
[1]  -3,5289989 -0,6730484 0,0056094 0,6482582 3,0345093
> rezumat(x)
       Min. Primul Qu. Media media a 3-a Qu. Max.
-3,529000 -0,672700       0,005609 -0,025180       0,648200      3,035000
> hist(x, freq=F) # Ieșirea histogramei (doar) în Figura 5.1 .
```

> curve(dnorm(x), add=T) # *Ieșirea curbei (doar) în* Figura 5.1 . >

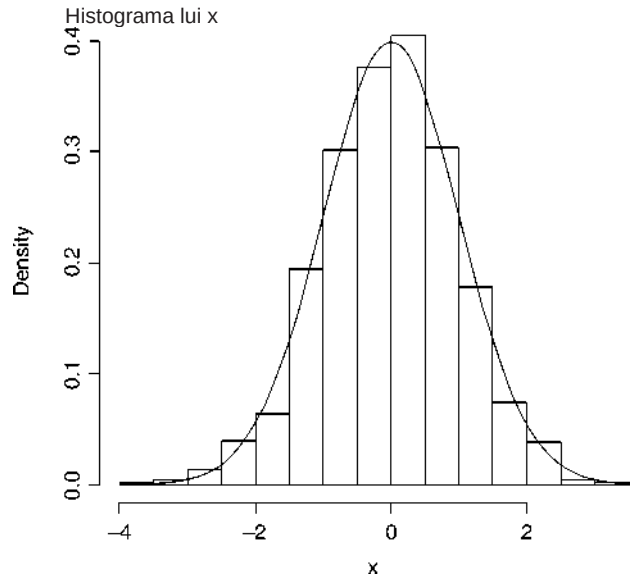


FIGURA 5.1 Distribuția a 1.000 de numere extrase aleatoriu din distribuția normală standard: folosind funcțiile hist() și curve().

INTERVALUL DE ÎNCREDERE (CI)

Exemplul 5.12 arată o interpretare probabilistică a intervalelor de încredere (IC).

■ **Exemplul 5.12:** Interpretarea probabilistică a CI

Un om de știință biomedical, care investiga nivelul mediu al unei enzime E într-o anumită populație, a abordat un eșantion aleatoriu de 100 de indivizi, a determinat nivelul enzimei la fiecare subiect și a calculat o medie a probei de $\underline{x}$ = 25 de unități. Se știa că nivelurile de enzime sunt distribuite aproximativ normal cu o variație de 400 de unități. Cum poate fi calculat nivelul mediu al enzimei μ al populației?

Soluţie:
CI de aproximativ 95% pentru *m* este

$$\underline{x} \pm 2\, s_{\underline{x}} = 25 \pm 2\sqrt{(400/100)} = 25 \pm 4 = [21, 29]$$

La calcularea acestui CI, rețineți că intervalul conține:

(a) Estimarea punctuală a lui *m* ca centru.
(b) Factorul 2, care decurge din distribuția normală standard, care arată numărul de erori standard (SE) care se află în aproximativ 95% din valoarea posibilă a lui $\underline{x}$. Această valoare a lui *z* este *coeficientul de fiabilitate* .

(c) Componenta $s_{\underline{x}}$ este eroarea standard sau abaterea standard a distribuției de eșantionare a lui $\underline{x}$.

Astfel, în general, o estimare de interval este reprezentată ca

Estimator ± (Coeficient de fiabilitate) × (Eroare standard) (5.1-23)

Prin urmare, pentru o probă luată dintr-o distribuție normală cu varianță cunoscută, intervalul pentru μ poate fi scris ca

$$x \pm z_{(1-\alpha/2)}\sigma_{\bar{x}} \quad (5,1\text{-}24)$$

unde $z_{(1-\alpha/2)}$ este valoarea lui *z* la stânga căruia se află (1 - α⁄2), iar la dreapta căreia se află *α⁄2* a ariei de sub curba de distribuție .

INTERPRETARE BAZAT PE TEORIA PROBABILITĂȚII. În Exemplul 5.12, unde coeficientul de fiabilitate a avut valoarea de 2, se poate afirma că la eșantionarea aleatorie repetată, aproximativ 95% din intervalele construite prin ecuația (5.1-24) vor include media populației. Aceasta se bazează pe probabilitatea de apariție a diferitelor valori ale *lui x* .

Desemnând aria totală sub curba de distribuție a lui x care se află în afara intervalului m ± 2 s χ_{CA} α *și* aria din intervalul ca 1 - *α* , se poate generaliza această interpretare pentru a ajunge la interpretarea probabilistă a CI după cum urmează .

Prin eșantionare aleatorie repetată dintr-o populație distribuită normal cu un SD cunoscut, 100(1 - *α*)% din toate intervalele formei

$$x \pm z_{(1-\alpha\sqrt{2})} \quad (5,1\text{-}25)$$

va include în cele din urmă media populației *m* .

Mărimea (1 - *α*), care este 0,95 în acest caz, este **coeficientul de încredere** (sau **nivelul de încredere**), iar intervalul $\boldsymbol{x \pm z_{(1-\alpha/2)}\mathbf{s}_{x}}$ este *CI* pentru μ. Când (1 - *α*) = 0,95, intervalul se numește „IC 95% pentru *m* ”. În Exemplul 5.12 , cercetătorul biomedical poate afirma că are 95% încredere că media populației este între 21 și 29. Aceasta este interpretarea practică a ecuației (5.1-24) .

OBSERVAȚII PRIVIND CI

1. **O interpretare practică**
   Când eșantionarea este luată aleatoriu dintr-o populație distribuită normal cu un SD cunoscut, se poate fi 100(1 - *α*)% sigur că intervalul calculat $x \pm z_{(1-\alpha/2)}\sigma_{\bar{x}}$ *conține* media reală a populației *m* .

   Valoarea mai exactă a lui *z* este 1,96, în loc de 2, corespunzând unui coeficient de încredere de 0,95. Coeficienții de încredere utilizați în mod obișnuit sunt 0,90, 0,95 și 0,99, pentru care factorii de fiabilitate sunt 1,645, 1,96 și, respectiv, 2,58.
2. **Quantile, mediană și CI**
   Inversa funcției de distribuție cumulativă (cdf) se numește funcție **cuantilă .** P - **quantila** este valoarea astfel încât există o probabilitate *p* de a obține o valoare egală sau mai mică decât acea valoare. Mediana este definită ca cuantila de 50% **.**

   Pentru calculul CI se folosesc cuantile teoretice. Astfel, pentru *n* observații distribuite normal cu medie *m* și SD *s*, media *x* este în mod normal distribuită în jurul *m* cu SD $s/\sqrt{n}$.

   Un CI de 95% pentru *m* poate fi obținut ca

$$x + s/\sqrt{n} \times N_{0.025} \le m \le x + s/\sqrt{n} \times N_{0.975} \quad (5.1\text{-}26)$$

   unde $N_{0,025}$ este cuantila de 2,5% din distribuția normală.

   De exemplu: Pentru *s* = 10, într-un studiu în care au fost testați 6 subiecți de caz și a fost găsită o medie de 75 de unități, se pot calcula cuantilele și CI folosind următorul segment de cod R :

```
>  xbar <- 75
>  sigma <- 10
>  n <- 6
>  # Cuantile și calcul CI:
>  sem <- sigma/sqrt(n) # sem = eroarea standard a mediei
> sem
>  1] 4,082483
>  xbar + sem * qnorm(0,025)
[1] 66,99848
>  xbar + sem * qnorm(0,975)
[1]  83,00152
```

Prin urmare, un CI de 95% pentru *m* este [66,99848, 83,00152] sau [70, 83] aproximativ.

Note:

1. Dacă distribuția normală este simetrică, atunci $N_{0,025} = N_{0,975}$.
2. Se poate exprima CI ca $\underline{x} \pm (s/\sqrt{n}) N_{0,975}$.
3. Cuantila poate fi exprimată ca $\Phi^{-1}(0,975)$, unde Φ este cdf al distribuției normale, pnorm() .
4. O altă aplicație a cuantilelor este prin utilizarea diagramelor cuantile-versus-quantile (Q–Q), care pot fi utilizate pentru a testa dacă se poate presupune că un set de date provine dintr-o distribuție dată. Funcția R qqnorm() poate fi utilizată pentru astfel de afișaje grafice.

■ **Exemplul 5.13:** diagrame QQ de seturi de date generate aleator

Luați în considerare:

1. Un set de 10 numere generate aleatoriu și arată graficul său Q–Q.
2. Repetați procedura pentru un set de 10.000 de numere.

Soluţie:

Următoarele segmente de cod oferă diagramele Q–Q necesare:

```
>  x1 <- rnorm(10)
>  x1 # Pentru un set de date mic, tipăriți-l și aruncați o privire la el!
[1]  1,1076094 1,6250114 -0,2617057 -1,8640710 -0,7168488
[6]  0,5900851 -0,9575308 -1,3772980 0,8060826 -1,8170715
> qqnorm(x1) # ieșire: Figura 5.2 (A).
> x2 <- rnorm(10000)
> qqnorm(x2) # ieșire: Figura 5.2 (B).
```

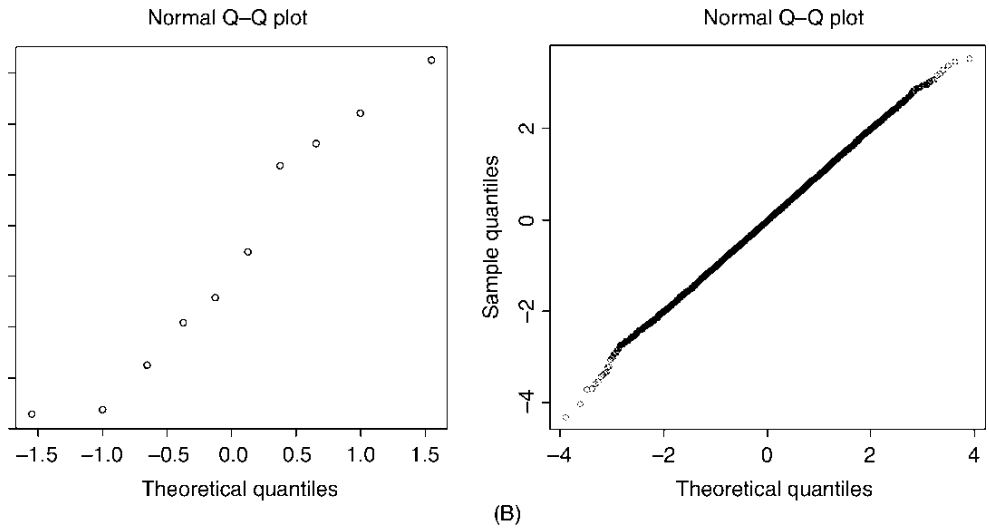


FIGURE 5.2 Q–Q plots of randomly sampled datasets from a normal distribution: (A) dataset of 10 points; (B) dataset of 10,000 points.

3. ***t* cu una și două probe pentru CI**

În analiza biostatistică a datelor reale, de cele mai multe ori, atât media populației *m* , cât și SD -*urile nu* sunt cunoscute. Din punct de vedere practic, sa considerat că o abordare rezonabilă ar fi realizarea unor estimări ale acestor parametri vitali, păstrând în același timp conceptul de distribuție normală pentru toate variabilele. Această abordare include:

- Folosind media eșantionului *x* ca estimare a mediei adevărate *m*
- Folosind SD- *uri* de probă pentru SD -*urile* adevărate
- Folosind distribuția *t* în locul distribuției normale
- În loc de valorile critice ale *lui* $z_{\alpha/2}$ într- o distribuție normală, folosind valorile critice mai mari ale valorilor $t_{\alpha/2}$ din distribuția *t*

Dacă o distribuție *t dată* este aproape de o distribuție normală (aceasta este ipoteza esențială), atunci distribuția variabilei *t* definită de

$$t = (\underline{x} - m)/(s/\sqrt{n}) \tag{5.1-27}$$

este o distribuție *t pentru eșantioanele de mărimea n* .

Această distribuție, cunoscută și ca distribuția *t a lui Student* , poate fi utilizată pentru a găsi valoarea critică $t_{\alpha/2}$. Numărul de grade de libertate (DF) este ($\boldsymbol{n}$ – 1), reprezentând numărul de mijloace *liber atribuibile n* subiecți de caz. În practică, este de așteptat ca $\boldsymbol{n} > 30$.

Au fost dezvoltate o serie de proceduri de testare a datelor. Unul dintre cele mai comune este **testul *t* cu un eșantion** , care presupune că seturile de date provin din distribuții normale. Pentru cazul cu un eșantion, datele $x_1, x_2, x_3, \ldots, x_n$ sunt presupuse a fi rezultate independente ale variabilelor aleatoare cu distribuție $N(ms^2)$); adică distribuția normală cu media *m* şi varianţa s^2. Ipoteza nulă de testat este $m = m_0$.

Se pot estima apoi parametrii *m* și *s* prin următoarele relații empirice:

$$m \approx \underline{x} \text{ și } s \approx s \tag{5.1-28}$$

În această abordare, conceptul esențial este că **eroarea standard a mediei** (SEM), care descrie variația mediei a *n* valori aleatoare cu media μ și varianța s^2, este

$$SEM = s/\sqrt{n} \tag{5,1-29}$$

Aceasta înseamnă că dacă testul ar fi repetat de mai multe ori și media calculată de fiecare dată, aceste medii ar forma o distribuție ceva mai îngustă decât cea a distribuției inițiale. Pentru datele distribuite normal, este de așteptat să existe o probabilitate de 95% ca datele să fie într-un interval de ($m \pm 2\,s$). Mai mult, ne-am aștepta ca $\underline{x}$ să fie în 2 SEM de el. În practică, se calculează:

$$t = (\underline{x} - m_0)/SEM \tag{5,1-30}$$

și apoi verifică pentru a vedea dacă acest parametru calculat se află într-o regiune de acceptare, în afara căreia *t* ar trebui să scadă cu probabilitate egală cu un nivel de semnificație predeterminat. Acest nivel este de obicei considerat ca fiind de 5%, corespunzând aproximativ intervalului [2, 2]. Valorile pentru regiunea de acceptare pot fi găsite ca cuantile în tabelele de distribuție *t* cu (n – 1) DF.

Criterii de acceptare :

- Dacă *t* se află în afara regiunii de acceptare , atunci ipoteza nulă este respinsă la nivelul de semnificație predeterminat.
- Se poate calcula , de asemenea, valoarea *p* (probabilitatea de a obține o valoare mai mare decât valoarea *t observată*) și, dacă valoarea *p* este mai mică decât nivelul de semnificație, se respinge ipoteza nulă.
- Un **test *t* unilateral** este efectuat dacă există informații suplimentare, astfel încât datele suplimentare ar determina doar *m* să depășească m_0. Într-un astfel de caz, se poate decide apoi ca ipoteza nulă să fie respinsă numai dacă *t* ar trebui să se încadreze în coada superioară a distribuției.

NIVELURI DE SEMNIFICAȚIE

În testele biostatistice, trebuie să decidem care valori intră în regiunea de respingere și care intră în regiunea de nonrejectare (aceasta din urmă nu este neapărat aceeași cu regiunea de acceptare). Această decizie se bazează pe **nivelul de semnificație preferat** α , care specifică aria de sub curbele de distribuție a statisticii de testare care este deasupra valorilor pe axa orizontală ca constituind **regiunea de respingere** .

Testele de ipoteză sunt numite și **teste de semnificație** , iar o valoare calculată a statisticii testului care se încadrează în regiunea de respingere este numită **semnificativă** . Astfel, nivelul de semnificație α este probabilitatea de a respinge o ipoteză nulă adevărată.

De obicei, valorile mici ale lui α sunt alese pentru a reduce probabilitatea de respingere a unei ipoteze nule adevărate. Frecvent, valorile alese ale lui α sunt 0,01, 0,05 și 0,10, 0,05 fiind cel mai frecvent.

Revedeți discuția din Capitolul 2 despre conceptul de semnificație biostatistică și luați în considerare următoarele exemple privind calculele nivelurilor de semnificație în analiza biostatistică.

> **Exemplul 5.14:** Niveluri de semnificație pentru teste cu două eșantioane pentru populații normale (Verzani, 2005)

Primul medicament antiretroviral susținut de Food and Drug Administration (FDA) utilizat în îngrijirea subiecților infectați cu HIV a fost azidotimidina (AZT). Doza normală a fost de 300 mg de două ori pe zi. Cercetătorii din domeniul sănătății știau că dozele mai mari au provocat adesea mai multe efecte secundare adverse, dar s-au întrebat dacă sunt și mai eficiente. În 1900, un studiu epidemiologic a comparat doze diferite la niveluri de 300, 600 și 1500 mg. Investigația a constatat că dozele mai mari au avut niveluri ridicate de toxicitate; ei au remarcat, de asemenea, ceva mai neașteptat, că dozele mai mici ar putea fi la fel de eficiente.

Antigenul p24 stimulează răspunsurile imune. Măsurarea nivelurilor p24 pentru grupurile de 300 mg și 600 mg este dată de datele simulate din Tabelul 5.6 .

Soluţie:

În primul rând, se stabilește o ipoteză nulă. Apoi este testat la niveluri de semnificație specificate.

Fie μ_{300} media grupului de 300 mg și μ_{600} să fie media grupului de 600 mg.

Se pot testa ipotezele nule ($H_0 : \mu_{300} = \mu_{600}$, $H_A : \mu_{300} \neq \mu_{600}$) cu un test *t* . Înainte de testare, verificați pentru a vedea dacă ipoteza unei variații comune și a unei normalități este adecvată prin inspectarea a două diagrame de densitate.

Următorul segment de cod R va îndeplini aceste sarcini:

```
>  # Introducerea celor două seturi de date: ×300 și ×600:
>  × 300 <- c(284, 279, 289, 292, 287, 295, 285, 279, 306, 298)
>  × 600 <- c(298, 307, 297, 279, 291.335, 299, 300, 306, 291)
>  plot(density(x300)) # Ieșire: Figura 5.3 (A).
>  lines(density(x600), lty=2) # Ieșire: Figura 5.3 (B).
>  t.test(x300, x600, var.equal=TRUE) # Acum, faceți testul t, rezultând:
```

Două eșantion de date t-test: x300 și x600 t = -2,034, df = 18, valoarea p = 0,05696
ipoteză alternativă: diferența adevărată în medii nu este egală cu 0 95 la sută
interval de încredere: -22,1584072 0,3584072
estimări ale eșantionului: media lui x media lui y 289,4 300,3

TABELUL 5.6 Nivelurile de p24 în miligrame pentru două populații de tratament

DOZARE	p24 NIVEL									
300 mg	284	279	289	292	287	295	285	279	306	298
600 mg	298	307	297	279	291	335	299	300	306	291

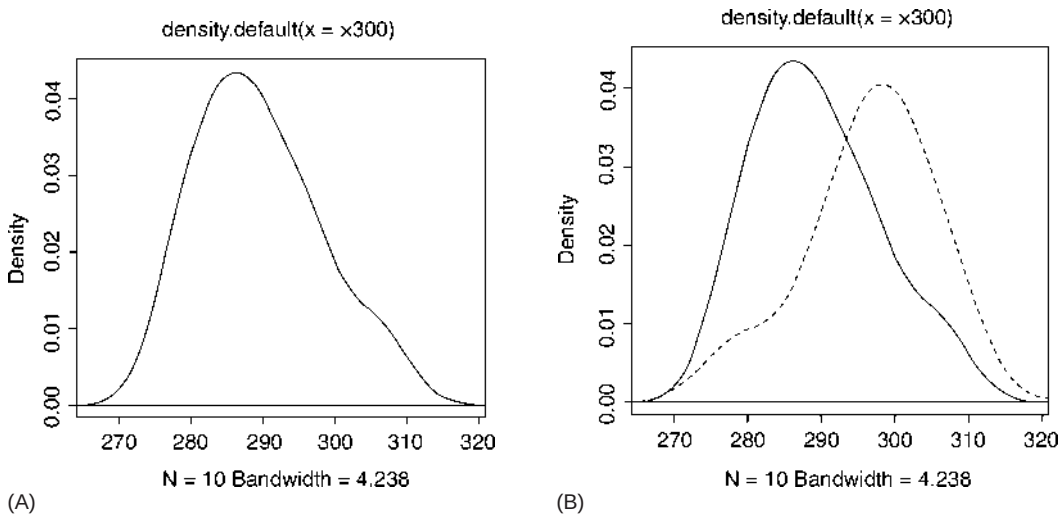


FIGURE 5.3 Density plots for comparing variances and shapes of the 300-mg dosage (solid) and the 600-mg dosage (dashed).

> *# Repetați testul t, fără a presupune aceleași variații*
> t.test(x300, x600)

Testul t Welch Two Sample
date: x300 și x600
t = -2,034, df = 14,509, valoarea p = 0,06065
Ipoteza alternativă: adevărata diferență de medii nu este egală cu 0
Interval de încredere de 95%:
-22,3557409 0,5557409
estimări ale eșantionului:
media lui x media lui y
289,4 300,3

Observatii:

1. Pentru testul de varianță egală, valoarea *p este .05696 pentru testul t* cu două fețe . Aceasta implică o posibilă diferență a valorilor medii, dar nu este semnificativă statistic la nivelul de semnificație de 0,05. Se poate concluziona că acest set de date este în concordanță cu ipoteza că nu există nicio diferență medie.
2. Statistica testului este t = -2,034.
3. Pentru cazul fără ipoteza variațiilor egale, se obține aceeași statistică de test (ca și cazul varianței egale): t = -2,034 . Acest acord este în concordanță cu modelul teoretic, după cum urmează:

   Din punct de vedere algebric, dacă două eșantioane sunt independente cu populații distribuite normal , așa cum X și Y estimează m_x și respectiv m_y , valoarea lui ($\underline{X} - \underline{Y}$) poate fi considerată o estimare bună pentru ($m_{x).} - al_{meu}$). Se poate folosi această ipoteză pentru a forma o statistică de testare.

   Ambele medii de eșantionare au distribuții de eșantionare distribuite în mod normal și, prin urmare, ar fi o statistică de test naturală T

$$T = \{(\underline{X} - \underline{Y}) \; E \; (\underline{X} - \underline{Y} \mid H_0)\}/\mathrm{SE}(\underline{X} - \underline{Y} \mid H_0) \qquad (5.1\text{-}31)$$

Sub H_0, valoarea așteptată a diferenței este zero.

SE poate fi găsită din formula pentru SD, care se bazează pe independența eșantioanelor:

$$SD(\underline{X} - \underline{Y} \mid H_0) = \sqrt{\{(s_x^2/n_x) + (s_y^2/n_y)\}} \quad (5.1\text{-}32)$$

Ca și în cazul calculului pentru CI, estimarea utilizată pentru variațiile populației depinde de ipoteza variațiilor egale.

Dacă se face ipoteza unei variații egale, atunci ambele seturi de date sunt reunite pentru estimare

$$s = s_x = s_y \quad (5,1\text{-}33)$$

folosind

$$s_p = \sqrt{[\{(n_x 1) s_x^2 + (n_y - 1) s_y^2\}/(n_x + n_y 2)]} \quad (5.1\text{-}34)$$

SE este dat de

$$SE(\underline{X} - \underline{Y}) = s_p \sqrt{\{(1/n_x) + (1/n_y)\}} \quad (5.1\text{-}35)$$

În aceste condiții, T are o distribuție t *cu* $(n - 2)$ DF.

Dacă se presupune că variațiile populației *nu* sunt egale, atunci se poate estima $s_{x\ cu}$ $s_{x\ și}$ $s_{y\ cu}$ $s_{y\ pentru}$ a obține

$$SE(m_x - m_y) = \sqrt{\{(s_x^2/n_x) + (s_y^2/n_y)\}} \quad (5.1\text{-}36)$$

Astfel, cu ipoteza variațiilor egale, ecuațiile (5.1-35) și (5.1-36) arată că atunci când cele două eșantioane au aceeași dimensiune, SE-urile lor sunt identice din punct de vedere matematic.

4. Valoarea p a crescut de la .05696 la .06065, deoarece DF-urile au scăzut de la 18 la 14.509.

Multe teste suplimentare, cum ar fi Wilcoxon {stats} și pairwise.wilcox. test {stats} , sunt disponibile pe site-ul CRAN la solicitarea de către > ??wilc o xon

BUNĂTATEA POTRĂRII

Un test **de bunăstare a potrivirii** (**GoF**), utilizat pe un set de date dat, evaluează statistic ipoteza conform căreia distribuția de frecvență a datelor se conformează sau „se potrivește" cu o anumită distribuție. În biostatistică, GoF-ul unui model arată cât de bine se potrivește modelul unui set de observații. Măsurile rezumative ale GoF constau, de obicei, în discrepanța dintre valorile *observate* și valorile *așteptate* conform modelului. Astfel de măsuri pot fi utilizate în testarea ipotezelor statistice în mai multe scopuri, inclusiv:

- Pentru a testa normalitatea setului de date sau a reziduurilor
- Pentru a testa dacă frecvențele rezultatelor urmează o distribuție specificată (vezi discuția despre testul chi-pătrat al lui Pearson mai târziu în această subsecțiune)
- Pentru a testa dacă două mostre sunt extrase din distribuții identice (a se vedea discuția despre testul Kolmogorov–Smirnov [KS] mai târziu în această secțiune)
- În analiza varianței (ANOVA), pentru a testa dacă una dintre componentele în care este împărțită varianța poate fi o sumă de pătrate care nu se potrivește (printre altele)

În mod obișnuit, în investigațiile epidemiologice, se dorește să se verifice dacă un eșantion de valori observate ale unor variabile specifice este sau nu compatibil cu ipoteza că eșantionul a fost extras dintr-o populație de valori distribuită normal. Această procedură de testare constă în general în punerea datelor observate în intervale de clasă care se exclud reciproc și în notarea frecvenței de apariție a valorilor în fiecare clasă. Apoi, folosind cunoștințele despre distribuțiile normale, se determină frecvențele pentru aceste clase la care ar putea fi așteptate dacă eșantionul ar fi provenit dintr-o distribuție normală. Dacă dezacordul este de așa natură

încât s-ar fi putut produce din cauza întâmplării, se poate concluziona că eșantionul observat poate să provină într-adevăr dintr-o distribuție normală. În mod similar, testele GoF pot fi efectuate în investigații în care distribuția ipotetică este distribuția binomială, distribuția multinomială, distribuția Poisson și așa mai departe.

Din teoria probabilității, au fost dezvoltate o serie de metodologii pentru testarea GoF. Aceste metodologii sunt prezentate aici și urmate cu exemple pentru a ilustra detaliile testării ipotezelor GoF folosind aceste metode în mediul R.

DISTRIBUȚIA CHI-SQUARE PEARSON PENTRU TESTUL GOF. Pentru populațiile care sunt - distribuite normal, cu varianță s^2, selectați aleatoriu eșantioane independente de dimensiunea n și calculați varianța eșantionului s^2, unde

$$s^2 = [\sum x_i^2 (\sum x_i)^2 / n]/(n - 1) \tag{5.1-37}$$

pentru fiecare probă. Statistica eșantionului c^2, unde

$$c^2 = (n - 1) s^2 / s^2 \tag{5,1-38}$$

are o distribuție numită **distribuție chi-pătrat** , are următoarele proprietăți speciale:

- Această distribuție se bazează pe numărul de DF, care este de obicei luat ca ($n - 1$) cu excepția cazului în care există condiții speciale.
- Această distribuție nu este simetrică, iar valorile nu pot fi negative; poate fi doar zero sau pozitiv.
- Pe măsură ce DF crește, această distribuție se apropie de o distribuție normală.

Statistica **testului** pentru testul chi-pătrat este

$$\boldsymbol{X}^2 = \sum [(O_i - E_i)^2 / E_i] \tag{5.1-39}$$

unde O_i este frecvența observată pentru clasa a i -a a variabilei de interes și E_i este frecvența așteptată pentru clasa a i -a a variabilei de interes (dacă ipotezele nule H_0 au fost adevărate).

Cantitatea $\boldsymbol{X}^2$ este o măsură a măsurii în care, în orice situație, perechile de frecvențe observate și așteptate sunt de acord. Natura lui $\boldsymbol{X}^2$ este ca:

(a) Când există un **acord *perfect*** între frecvențele observate și cele așteptate , acesta este ***zero*** și H_0 trebuie acceptat.
(b) Când există un **acord strâns** între frecvențele observate și cele așteptate, acesta este ***mic*** (și valoarea p este mare).
(c) Când există un **acord slab** între frecvențele observate și cele așteptate, acesta este ***mare*** (și valoarea p este mică), iar H_0 trebuie respins.

În general, valoarea calculată a *lui* X^2 este comparată cu valoarea tabelată a lui ($k - r$) DF, unde k este numărul de clase pentru care sunt disponibile frecvențele observate și așteptate și r este numărul de constrângeri sau restricții impuse comparației date. Astfel, se impune o restricție atunci când se forțează ca suma frecvențelor așteptate să fie egală cu suma frecvențelor observate. Restricții suplimentare sunt impuse pentru fiecare parametru care este estimat din eșantionul observat.

Decizia este atunci dacă se respinge ipoteza nulă H_0 dacă X^2 este mai mare sau egal cu c^2 tabelat pentru valoarea aleasă a lui α .

Tabele cu valorile critice ale c^2 distribuția sunt disponibile în texte standard și pe Internet.

TESTUL CHI-PĂTRATUL LUI PEARSON ÎN MEDIUL R. Funcția R chisq . test() , în pachetul CRAN statistics , efectuează teste de tabel de contingență chi-pătrat și teste GoF. Formatul său de

utilizare este chisq.test(x, y = NULL, corect = TRUE,
p = rep(1/lungime(x), lungime(x)), redimensionare.p = FALS,
simulare.p.valoare = FALS, B = 2000)
pentru care argumentele sunt:

x Un vector numeric sau o matrice. x și y pot fi, de asemenea, ambii factori.

y Un vector numeric; ignorat dacă x este o matrice. Dacă x este un factor, atunci y ar trebui să fie un factor de aceeași lungime.

corect O etichetă logică care indică dacă se aplică continuitatea corect la calculul statisticii de testare pentru tabele 2 × 2 : o jumătate se scade din toate |*OE*| diferențe. Nu se face nicio corecție dacă simulate.p.value = TRUE .

p Un vector de probabilități de aceeași lungime cu x . O eroare este dat dacă orice intrare a lui p este negativă.

rescale.p Un scalar logic; dacă TRUE , atunci p este redimensionat (dacă este necesar) la sumă la 1. Dacă rescale.p este FALSE și p nu se însumează la 1, este dată o eroare.

simulate.p.value O etichetă logică care indică dacă să se calculeze valorile *p* prin simularea Monte Carlo.

B Un număr întreg care specifică numărul de replici utilizate în Testul Monte Carlo.

Următorul exemplu ilustrează utilizarea funcției chisq.test() .

> **Exemplul 5.15:** Testul GoF pentru distribuția normală (Daniel, 2005 , Verzani, 2005)

Un grup de cercetare epidemiologică a colectat datele privind ocuparea pacienților internați pe 250 de spitale din SUA pe o perioadă de 12 luni și a raportat raportul dintre recensământul zilnic și numărul de paturi menținute. Datele eșantionului au fost exprimate în termeni de distribuție a procentelor, așa cum se arată în Tabelul 5.7 .

Arătați dacă aceste date oferă dovezi rezonabile suficiente pentru a demonstra că eșantionul a provenit sau nu dintr-o populație distribuită normal.

TABELUL 5.7 Rezumatul rezultatelor studiului privind ocuparea spitalelor

RAPORTUL DE OCUPARE A PACIENȚILOR STAȚIONAȚI	NUMĂR DE SPITALARE
0,0 până la 39,9	16
40,0 până la 49,9	18
50,0 până la 59,9	22
60,0 până la 69,9	51
70,0 până la 79,9	62
80,0 până la 89,9	55
90,0 până la 99,9	22
100,0 până la 109,9	4

Total	250

Soluţie:

Începeți prin a presupune că setul de date disponibil pentru analiza biostatistică este un eșantion simplu, aleatoriu.

Ipoteza (H_0): În populația din care a fost prelevat acest eșantion, ratele de ocupare a pacienților internați sunt distribuite în mod normal.

Statistică de test: $X = \sum_{i=1 \text{ la } k} [(O_i E_i)^2 / E_i]$ (5.1-37)

Aceasta este statistica chi-pătrat, în care O_i și E_i sunt frecvențele observate și, respectiv, așteptate pentru a i -a clasă a setului de date.

Dacă H_0 este adevărată, atunci statistica testului este distribuită ca chi-pătrat cu ($k - r$) DF.

Regula de decizie: H_0 va fi respins dacă valoarea calculată a lui X^2 este egală sau mai mare decât valoarea critică a chi-pătratului sau valoarea p este foarte mică.

Calcule: Deoarece media și varianța distribuției ipotetice nu sunt cunoscute, datele eșantionului vor fi utilizate pentru a estima acești parametri. Estimările acestor parametri sunt necesare pentru a calcula frecvența la care ar fi de așteptat în fiecare clasă atunci când ipoteza este adevărată. Pentru fiecare dintre cele 8 clase, mediana fiecăreia va fi folosită pentru a reprezenta clasa; medianele celor 8 clase sunt 20, 45, 55, 65, 75, 85, 95 și 105.

Următoarele segmente de cod R sunt utilizate pentru a efectua calculele:

```
>  x <- c(16, 18, 22, 51,62, 55, 22, 4) # Introducerea frecvențelor
>  x # Verificare
[1]      16      18      22      51      62      55      22       4
>  pl <- c(20, 45, 55, 65, 75, 85, 95, 105) # Introducerea datelor
>  p1 # Verificare
[1]      20      45      55      65      75      85      95     105
>  p <- p1/sum(p1) # Normalizare
>  p # Verificare
[1] 0,03669725 0,08256881 0,10091743 0,11926606
[5] 0,13761468 0,15596330 0,17431193 0,19266055
>  n <- sum(x) # Însumând frecvențele
>  n # Verificare
[1] 250
>  chi2 <- sum( (x - n*p)^2 / (n*p) ) # Calcularea X² statistică
>  chi2 # Ieșirea X² statistică
[1] 100,7709
> pchisq(chi2, df = 8 - 1, lower.tail = F) # Ieșirea valorii p
[1] 7.476817e-19
>
>  # Un calcul alternativ:
>  test chisq.(x, p=p)
   Testul chi-pătrat pentru date probabilităților date: x
X-pătrat = 100,7709, df = 7, valoare p < 2,2e-16
```

Decizie și concluzie biostatistică: Probabilitatea de a obține o valoare a X^2 pentru a permite ipoteza H_0 să fie adevărată este mică: $p \approx 0$. Astfel, setul de date este puțin probabil să fie distribuit în mod normal și ar trebui să caute o altă explicație.

TESTUL CHI-PĂTRAT DE INDEPENDENȚĂ PENTRU GOF. Capitolul 2 a introdus utilizarea unui tabel de contingență bidirecțional pentru a clarifica relația dintre variabile. În special, cineva poate fi preocupat dacă nivelurile unei variabile afectează distribuția celeilalte variabile. Adesea apare întrebarea dacă acestea sunt variabile aleatoare independente. Ca o ilustrare, luați în considerare următoarea situație (Verzani, 2005) .

În California, un studiu privind utilizarea centurilor de siguranță a investigat relația dintre utilizarea centurii de siguranță de către un părinte și de către un copil. Rezultatul sondajului este prezentat în Tabelul 5.8 .

În timp ce datele arată unele diferențe dramatice și corelații importante între relații și acțiuni, din punct de vedere biostatistic, este nevoie de un test de semnificație - începând cu un mod de probabilitate, statistica de testare asociată și ipotezele nule și alternative care urmează să fie testate. ajunge la o decizie și obține un concept GoF al modelului. O astfel de abordare poate fi aplicată oricărei situații similare de decizie în patru direcții.

Modelul probabilității: Modelul de eșantionare este că fiecare mașină urmează o probabilitate dată care este înregistrată într-o anumită celulă. Aceste probabilități nu se modifică, iar rezultatul uneia nu afectează distribuția altuia; astfel, ele formează o secvență iid. Luați în considerare un model multinomial pentru setul de date.

TABELUL 5.8 Studiu privind utilizarea centurilor de siguranță auto în California

	COPIL	
MAMĂ	CUCAT	DESCACHAT
Cataramă	56	8
Desfăcut	2	16

Lasă

n_r = numărul de rânduri din tabel (numărul de niveluri ale variabilei rând) n_c = numărul de coloane din tabel

X_{ij} = O variabilă aleatorie care înregistrează frecvența celulei (i,j).

p_{ij} = Probabilitatea celulei pentru *al* i -lea rând și j -a coloană

Notați probabilitățile marginale cu p_i^r și p_j^c, unde:

$$P\Gamma = P_{i1} + P_{i2} + \text{L} + P_{in}, \tag{5,1-38}$$

și

$$P_c = P_{j1} + P_{j2} + \text{L} + P_{jn_t} \tag{5.1-39}$$

Ipoteza nulă: variabilele coloanei sunt independente de variabilele rând. adica

$$H_0 : pi = p_i^r p^c \tag{5,1-40}$$

Cu alte cuvinte, ipotezele sunt:

H_0 : Variabilele sunt independente.

H_A : Variabilele nu sunt independente.

După ce valorile p_i sunt estimate astfel încât să se calculeze numărul „așteptat”, statisticile c^2

$$c^2 = \sum [(\textit{observat} - \textit{așteptat})^2 / \textit{așteptat}] \tag{5.1-41}$$

poate fi folosit. Datele sunt folosite pentru a estima probabilitățile marginale (a se vedea secțiunea 5.1), iar ipoteza independenței este utilizată pentru a estima p_i .

Pentru acest exemplu, probabilitățile marginale sunt obținute prin distribuția

marginală a datelor, rezultând în Tabelul 5.9 .

Din Tabelul 5.9 , este ușor de observat că:

- Estimarea pentru $p^{r}{}_{1} = P$ (Părintele are cataramă) = $p^{r}{}_{1} = 64/82$
- Estimarea pentru $p^{r}{}_{2} = P$ (Părintele este dezlegat) = $p^{r}{}_{2} = 18/82$
- Estimarea pentru $p^{c}{}_{1} = P$ (Copilul are cataramă) = $p^{c}{}_{1} = 58/82$
- Estimarea pentru $p^{c}{}_{2} = P$ (Copilul este desfacut) = $p^{c}{}_{2} = 24/82$

După calcularea acestor estimări, se poate folosi ipoteza nulă H_0 pentru a găsi estimarea:

$$\underline{p}_{ij} = \underline{p}_{i}{}^{r}\, \underline{p}_{j}{}^{c} \qquad (5,1\text{-}40)$$

Pentru datele disponibile, estimările rezultate sunt obținute așa cum se arată în Tabelul 5.10 .

TABELUL 5.9 Distribuția marginală pentru studiul utilizării centurii de siguranță în California

	COPIL		
MAMĂ	CUCAT	DESCACHAT	MARGINAL
Cataramă	56	8	64
Desfăcut	2	16	18
Marginal	58	24	82

TABELUL 5.10 Probabilități marginale estimate pentru studiul utilizării centurii de siguranță auto în California

	COPIL		
MAMĂ	CUCAT	DESCACHAT	MARGINAL
Cataramă	(64/82)(58/82)	(64/82)(24/82)	(64/82)
Desfăcut	(18/82)(58/82)	(18/82)(24/82)	(18/82)
Marginal	58/82	24/82	82/82

Folosind acest tabel, se pot calcula sumele așteptate în celula ij cu $n\,\underline{p}_{ij}$, adică e_{ij} , care este scris ca $R_i C_j / n$, unde R_i este suma rândurilor și C_j suma coloanei :

$$e_{ij} = R_i C_j / n \qquad (5,1\text{-}42)$$

iar c^2 statistica poate fi acum exprimată ca

$$x^2 = \sum_{i=1 \text{ toná } r} \sum_{i=1 \text{ toná } c} \lceil \lfloor \left(Y_{ij} - n\,\underline{p}_{ij} \right)^2 n\,\underline{p}_{ij} \rceil \rfloor \qquad (5.1\text{-}43)$$

Prin ipoteza datelor multinomiale și independența variabilelor, distribuția de eșantionare a *lui* c^2 va fi distribuția chi-pătrat cu $(n_r - 1)(n_c - 1)$ DF, deoarece scăzând 1 DF din $n_r \cdot n_c - 1$ pentru fiecare parametru estimat, și întrucât există $n_r - 1 + n_c + 1$ parametri, valoarea pentru DF este

$$n_r \cdot n_c - 1 - (n_r - 1 + n_c + 1) = n_r \cdot n_c - n_r - n_c + 1 = (n_r - 1)(n_c - 1)$$

În mediul R , se poate urma următoarea procedură pentru un test chi-pătrat pentru independența a două variabile categoriale:

- Când datele sunt rezumate într- o matrice sau tabel în variabila x , testul este realizat de funcția de testare chisq.test().
- Când datele sunt rezumate și stocate în două variabile *x* și *y* unde intrările *i* se potrivesc, funcția de testare este chisq.test(x, y).
- Datele sunt mai întâi rezumate folosind table() , ca în chisq.test(table(x, y)).

Ipotezele nule și alternative nu sunt specificate, deoarece sunt aceleași pentru fiecare test. Când numărul așteptat din unele celule este prea mic pentru a utiliza distribuția chi-pătrat pentru a reprezenta distribuția de eșantionare a lui c^2, adăugarea argumentului simulate.p.value=TRUE va returna o estimare a valorii *p* folosind o simulare Monte Carlo .

Următorul exemplu ilustrează utilizarea lui R pentru această problemă.

- **Exemplul 5.16:** Testul GoF folosind testul Pearson χ^2 al independenței (Verzani, 2005)

Folosind datele din Tabelul 5.8 , sondajul privind utilizarea centurii de siguranță auto în California, efectuează un test GoF folosind R .

Soluţie:

Următoarele segmente de cod R pot fi utilizate pentru a crea mai întâi un tabel utilizând funcția rbind() pentru a combina rânduri (vezi Capitolul 3 despre calcule în vectori și grafice simple), apoi folosind funcția chisq.test() pentru a face Test GoF:

```
> seatbelt.usage.in.california <- rbind( c(56,8), c(2,16) )
> seatbelt.usage.in.california # ieșire:
     [,1] [,2]
[1,] 56    8
[2,] 2     16
> chisq.test(seatbelt.usage.in.california) # ieșire:
```

Testul Chi-pătrat al lui Pearson cu datele de corecție a continuității lui Yates:
seatbelt.usage.in.california
X-pătrat = 35,9953, df = 1, valoarea p = 1,978e $^{-09}$

Decizie și concluzie biostatistică: Valoarea *p* calculată , 1.978e $^{-09}$, este mică, nu este semnificativă și, prin urmare, este în concordanță cu observația că cele două variabile *nu* sunt independente.

TESTUL CHI-PĂTRAT DE HOMOGENEITATE PENTRU GoF. În investigarea eficacității unui tratament medicamentos, epidemiologii de cercetare sunt adesea chemați să evalueze un studiu clinic în care fiecare subiect de caz este alocat aleatoriu unuia dintre cele două grupuri: fie un grup de tratament, fie un grup placebo. Pentru a analiza biostatistic rezultatele, sunt disponibile următoarele abordări:

(A) Dacă rezultatele sunt înregistrate numeric, se poate utiliza un test *t* pentru a testa dacă diferențele dintre mediile eșantionului sunt semnificative.
(B) Dacă rezultatele sunt notate categoric, următoarea procedură arată că statistica c^2 poate fi utilizată pentru a verifica dacă distribuțiile rezultatelor sunt aceleași. Un studiu de cercetare efectuat la Centrul Medical al Universității Stanford (SUMC) este folosit aici pentru a ilustra această abordare (Verzani, 2005) .

SUMC a investigat dacă antidepresivul Celexa poate fi eficient în modificarea -comportamentului compulsiv la cumpărături. În acel studiu, 24 de subiecți de caz (care erau cunoscuți cumpărători compulsivi) au participat: 12 au primit o doză zilnică de Celexa timp de 7 zile consecutive, iar 12 au primit un placebo. La sfârșitul acestui tratament, toți subiecții au fost chestionați pentru a evalua dacă dorința lor de a face cumpărături a fost redusă. Din raportul preliminar au fost elaborate date simulate; sunt prezentate în Tabelul 5.11 .

Pentru a formula această investigație ca test de semnificație, se pot folosi următoarele ipoteze:

H_0: Cele două distribuții sunt aceleași.
H_A: Cele două distribuții sunt diferite.

iar c^2 pot fi folosite statistici. Deoarece cantitățile așteptate nu sunt pe deplin specificate în H_0, acea cantitate trebuie determinată.

TABELUL 5.11 Date privind efectul Celexa pentru reducerea cumpărăturilor compulsive

	MULT MAI RĂU	MAI RĂU	ACELEAŞI	MULT Imbunatatit	FOARTE MULT ÎMBUNĂTĂTIT
Celexa	0	2	2	5	2
Placebo	0	2	8	2	0

Procedura de testare este după cum urmează. În tabelul de date, lasă

> Variabila aleatorie să fie variabila **coloană** și
> Categoria care descompune datele este variabila **rând** .
> Pentru rândul i al tabelului, fie p_{ij} probabilitatea ca variabila aleatoare (rezultatul studiului) să se afle la nivelul j al variabilei aleatoare.

Putem apoi reformula ipotezele ca

$$H_0: p_{ij} = p_j \text{ pentru toate rândurile } i$$
$$H_A: p_{ij} \neq p_j \text{ pentru unele } i, j$$

Fie n_i numărul de numărări din fiecare rând; atunci cantitatea așteptată în celula (i, j) sub H_0 ar trebui să fie $n_i p_j$. Valoarea lui p_j în H_0 trebuie estimată deoarece nu este specificată. În ipoteza H_0, toate datele din coloana j -a sunt binomiale cu n și p_j; astfel, un estimator pentru p_j ar fi suma coloanei împărțită la n; adică C_j / n.

În aceste circumstanțe, numărul așteptat în celula (i, j) ar fi dat de

$$e_{ij} = n_i \underline{p}_j = R_i C_j / n \qquad (5.1\text{-}44)$$

Aceasta este exact aceeași formulă ca și c^2 test de independență, Ecuația (5.1-42) .

În ciuda diferențelor dintre ipoteze, statistica de test și distribuția ei de eșantionare în ipoteza nulă H_0 sunt **aceleași** cu testul de independență; prin urmare, testele de semnificație chi-pătrat ale omogenității și independenței sunt identice în practică și implementare.

TESTE GoF PENTRU DISTRIBUȚII CONTINUE. Amintiți-vă că atunci când se testează dacă datele eșantionate au fost luate dintr-o distribuție normală, se face o histogramă sau un grafic cantitativ al datelor, apoi se inspectează vizual rezultatul. Pentru un set de date continuu eșantionat, se poate compara, folosind un test de semnificație, setul de date cu

unul teoretic. Aceasta este abordarea testelor GoF pentru distribuții continue.

Pentru datele categorice, s-a demonstrat că poate fi aplicat testul chi-pătrat. Se poate extinde această tehnică prin „binning": la fel cum se face atunci când construim histograme , alegem niște bins și numărăm numărul de puncte de date din fiecare bin. În acest fel, datele pot fi considerate categorice, iar testul poate fi utilizat pentru GoF.

Totuși, în practică, s-a constatat că, pentru distribuțiile continue, o abordare care îmbunătățește rezultatele este **testul KS** : Fie X_1, X_2, X_3, …, X_i, …, X_n un eșantion aleatoriu din o distribuție continuă.

Fie $f(x)$ funcția de densitate a distribuției de probabilitate (pdf), iar X o altă variabilă aleatoare cu această densitate.

Cdf pentru X este $F(x) = P(X \leq x)$ sau aria din stânga lui x sub curba de densitate a lui X .

Cdf poate fi definit în mod similar atunci când X este discret: Se calculează din pdf prin însumarea: $P(X \leq x) = \sum_{y \leq x} f(y)$.

Pentru un eșantion X_1, X_2, X_3, …, X_i, …, X_n, **distribuția empirică** este distribuția obținută prin eșantionare din punctele de date. „Probabilitatea ca un număr selectat aleatoriu dintr-un eșantion să fie ≤ x" este numărul de puncte de date din eșantion ≤ (x / n). Folosind notația $F_n(x)$ pentru această probabilitate:

$$F_n(x) = \#\{ i : X_i \leq x \} / n \tag{5.1-45}$$

Aici, $F_n(x)$ este *funcția de distribuție cumulativă empirică* (ecdf).

În mediul R , $F_n(x)$ poate fi reprezentat folosind funcția ecdf() , în pachetul statistics , similar cu funcția density() , pentru care valoarea returnată poate fi reprezentată într-o nouă cifră folosind funcția plot( ) , sau poate fi adăugat la graficul existent folosind funcția lines() . Următorul exemplu ilustrează utilizarea lui R pentru această aplicație.

> **Exemplul 5.17:** Compararea unui eșantion de 15 puncte dintr-o populație distribuită normal cu distribuția teoretică, arătând atât densitățile eșantionului, cât și teoretice și cdfs.

Soluţie:
Pot fi utilizate următoarele comenzi R :

```
> x1 <- rnorm(15)
> x1
```

[1] **-0,63573743** **-0,14453713** **-0,57131893**
[4] **-1,76854783** **1,41941167** **-1,47808847**
[7] **0,99010104** **-0,06643542** **0,25797379**
[10] **-0,62145348** **-0,77645263** **0,31534275**
[13] **-0,96369367** **-1,15635521** **-1,08556058**

```
> plot(density(x1), main="Densities") # Trasarea densităților
> curve(dnorm(x), add=TRUE, lty=2) # Adaugă curbă normală
> # ieșire: figurile 5.4(A) și 5.4(B) .
> plot(ecdf(x1), main="Cdfs") # Trasarea valorilor punctului Cdf
> curve(pnorm(x), add=TRUE, lty=2) # Adaugă curba Cdf
> # ieșire: figurile 5.5(A) și 5.5(B) .
```

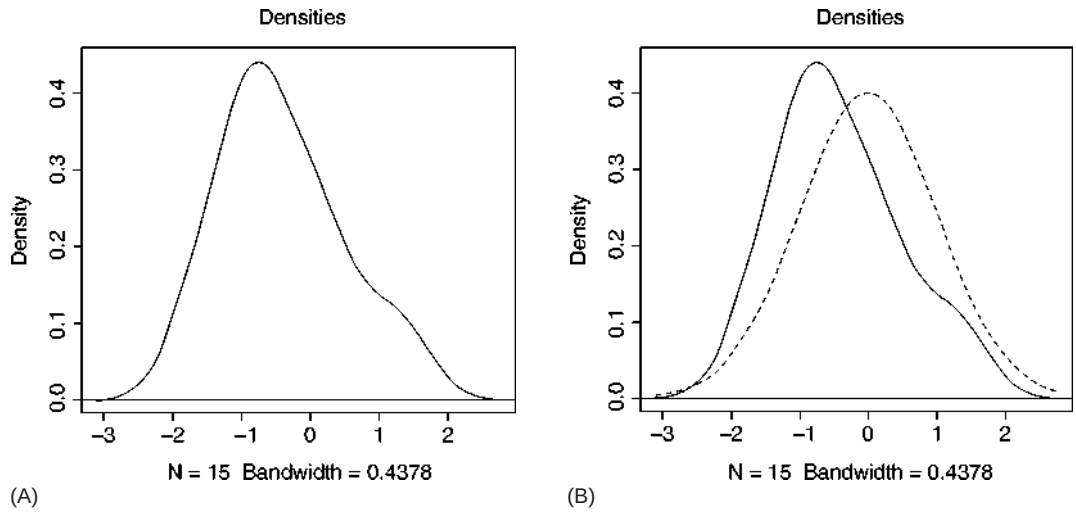


FIGURE 5.4 For a sample size of 15 from a normally distributed population: (A) sample density; (B) theoretical density.

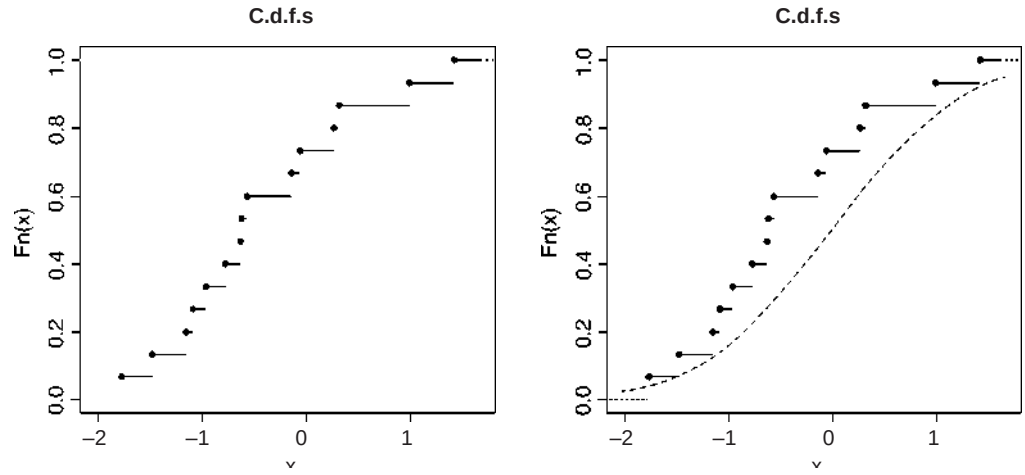


FIGURE 5.5 For a sample size of 15 from a normally distributed population: (A) estimated cdf; (B) theoretical cdf.

Observatii:

- Exemplul de mai sus arată că, dacă datele eșantionate sunt dintr -o populație cu distribuție normală, distribuția și cdf ale setului de date vor fi apropiate de distribuția normală și de cdf a distribuției normale.
- În mod similar, dacă datele provin dintr -o populație cu cdf F , ne putem aștepta ca F_n să fie aproape de F într-un fel. Ce înseamnă „închidere” în acest context?
- Având în vedere două funcții diferite ale *lui* x , separarea D dintre ele poate fi definită ca

$$D = \text{maxim de } |F_n(x) - F(x)| \tag{5.1-46}$$

Se pare că, cu singura presupunere că F este continuă, D are o distribuție de eșantionare cunoscută numită distribuție KS. Această proprietate permite construirea unui test de semnificație folosind statistica de test D .

Mai mult, un test similar poate fi efectuat pentru a compara două probe independente.

TESTUL K–M GoF. Fie $X_1, X_2, X_3, \ldots, X_i, \ldots, X_n$ o probă independentă și distribuită identic (iid) dintr-o distribuție continuă cu cdf $F(x)$.

Fie $F_n(x)$ cdf empiric.

Un test de semnificație de $H_0: F(x) = F_0(x)$, $H_A: F(x) \neq F_0(x)$ poate fi construit cu statistica de test D . Rețineți că valorile mari ale lui D indică sprijin pentru ipoteza alternativă.

În R , testul KS GoF poate fi implementat folosind funcția ks.test() , din pachetul CRAN stats , care realizează teste KS cu unul sau două eșantioane, în următorul format.

Din mediul R , comanda

```
> ?ks.test
```

furnizează următoarele informații despre documentația funcției.

ks.test() , în pachetul CRAN stat s [XIX]:
ks.test(x, y, … , alternativă = c(„două fețe", „mai puțin", „mai mare"), exact = NULL)

Argumente:

x Un vector numeric al valorilor datelor.
y Fie un vector numeric de valori de date, fie un șir de caractere nam ing un CDF sau un CDF real, cum ar fi pnorm . Numai CDf-urile continue sunt valide.
… Parametrii distribuției specificati (ca șir de caractere) de y .
alternativă Indică ipoteza alternativă și trebuie să fie una dintre „ două. cu laturi " (implicit), „ mai puțin " sau „ mai mare ". Puteți specifica doar litera inițială a valorii, dar numele argumentului trebuie dat în întregime. Consultați „Detalii" pentru semnificațiile valorilor posibile.
NULL exact sau o etichetă logică care indică dacă ar trebui calculată o valoare *p exactă.* Consultați „Detalii" pentru semnificația lui NULL . Acest lucru nu este disponibil în cazul cu două mostre pentru un test unilateral sau dacă sunt prezente legături.

Detalii:

- Dacă y este numeric, se efectuează un test cu două eșantioane al ipotezei nule că x și y au fost extrase din aceeași distribuție *continuă* .
- Alternativ, y poate fi un șir de caractere care denumește o funcție de distribuție continuă (cumulativă) sau o altă asemenea funcție. În acest caz, se efectuează un test cu un eșantion de nul pentru a vedea dacă funcția de distribuție care a generat x este distribuția y cu parametri specificați de … .
- Prezența legăturilor trebuie luată întotdeauna ca un avertisment, deoarece distribuțiile continue nu le generează. Dacă legăturile au apărut din rotunjire, testele pot fi aproximativ valide, dar chiar și cantități modeste de rotunjire pot avea un efect semnificativ asupra statisticii calculate.
- Valorile lipsă sunt omise în tăcere din x și (în cazul cu două eșantioane) y .
- Valorile posibile „ cu două fețe ", „ mai puțin " și „ mai mare " ale alternativei specifică ipoteza nulă conform căreia adevărata funcție de distribuție a lui x este egală, nu mai mică sau nu mai mare decât, funcția de distribuție ipotetică (un eșantion). caz) sau, respectiv, funcția de distribuție a lui y (caz cu două eșantioane). Aceasta este o comparație a cdfs, iar statistica de test este diferența maximă de valoare, statistica din alternativa „ mai mare " fiind *D^+ = max[F_x(u) - F_y(u)]* . Astfel, în cazul cu două eșantioane, alternativa = „mai mare" include distribuțiile pentru care x este stocastic *mai mic* decât y (cdf-ul lui x se află deasupra și, prin urmare, la stânga celui pentru y), spre deosebire de t.test sau wilcox .test .
- Valorile *p* exacte nu sunt disponibile pentru cazul cu două eșantioane dacă sunt unilaterale sau în prezența legăturilor. Dacă exact = NULL (implicit), se calculează o valoare *p exactă* dacă dimensiunea eșantionului este mai mică de 100 în cazul unui eșantion *și nu există legături;* iar dacă produsul dimensiunilor eșantionului este mai mic de 10.000 în cazul cu două eșantioane.

  În caz contrar, se folosesc distribuții asimptotice ale căror aproximări pot fi inexacte în eșantioane mici. În cazul unui eșantion, cu două fețe, valorile *p exacte sunt obținute așa*

[XIX] Pachetul CRAN {stats}: http://cran.r-project.org

cum este descris în Marsaglia, Tsang și Wang (2003) (dar fără a se folosi aproximarea opțională din coada dreaptă, deci aceasta poate fi lent pentru mici valorile *p*). Formula lui Birnbaum și Tingey (1951) este utilizată pentru cazul cu un eșantion, unilateral.

- Dacă se utilizează un test cu un singur eșantion, parametrii specificați la ... trebuie să fie prespecificați și nu estimați din date. Există o teorie mai rafinată a distribuției pentru testul KS cu parametri estimați (Durbin, 1973) , dar aceasta nu este implementată în ks.test .

Valoare:

O listă cu clasa „ htest ” având următoarele componente:

statistic Valoarea statisticii testului

p.valoare Valoarea *p* a testului

alternativă Un șir de caractere care descrie metoda alternativă de ipoteză Un șir de caractere care indică ce tip de test a fost efectuat

data.name Un șir de caractere care oferă numele (numele) datelor

Observatii:

Pentru aplicația curentă, funcția ks.test() poate fi utilizată ca ks.test(x, y), unde x și y stochează datele. Sunt disponibile teste cu unul și două mostre (Conover, 1971) .

Următorul exemplu ilustrează utilizarea lui R pentru această aplicație.

- **Exemplul 5.18:** Testul Shapiro-Wilk (SW) GoF pentru normalitate

Utilizați testul SW pentru GoF, în R , pentru a evalua următoarele seturi de date pentru normalitate:

(a) Un set de date de 1.000.000 de puncte selectate aleatoriu dintr-o distribuție normală cu medie 0 și SD 10

(b) Un set de date de 1.000 de puncte dintr-o distribuție uniformă continuă

Soluţie:

Pot fi utilizate următoarele segmente de cod R.

Pentru a genera setul de date dintr-o *distribuție normală* , se poate folosi funcția rnorm() :

```
> x <- rnorm(1000, medie = 0, sd = 10)
> shapiro.test(x)
```

Datele testului de normalitate Shapiro–Wilk: x

W = 0,9979, valoarea p = 0,2445

Observații: Valoarea *p* , 0,2445, nu este semnificativă statistic. Astfel, nu există dovezi că setul de date nu este distribuit în mod normal. H_0 nu poate fi respins.

Funcția de densitate de probabilitate a *distribuției uniforme continue* este

$$f(x) \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{for } x < a \text{ or } x > b \end{cases} \qquad (5.1\text{-}47)$$

Valorile lui $f(x)$ la cele două limite a și b nu afectează valorile integralelor lui $f(x)\,dx$ pe orice interval și nici ale $xf(x)\,dx$ sau vreun moment superior. Ele pot fi alese să fie zero, sau să fie $1/(b-a)$. Funcția de densitate de probabilitate și cdf-ul acestei distribuții sunt prezentate în figurile 5.6(A) și , respectiv, (B) .

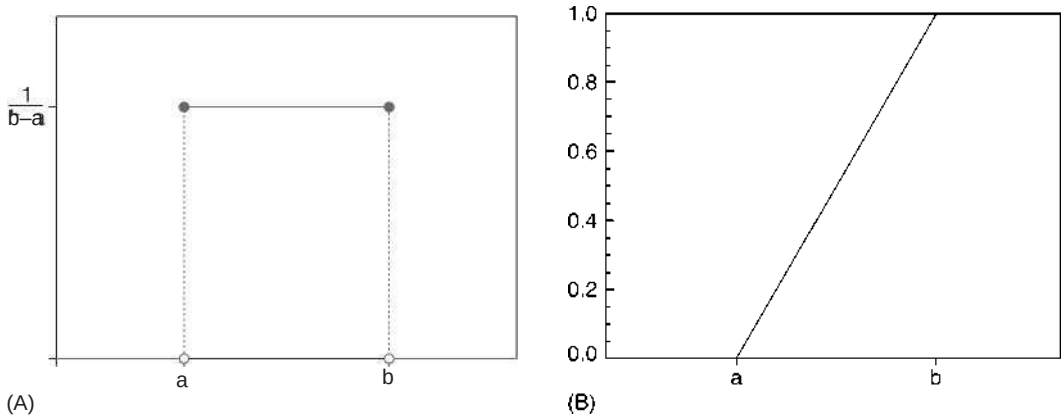


FIGURE 5.6 The uniform distribution function: (A) probability density function; (B) cumulative distribution function.

```
> z <- runif(1000) # Generarea a 1000 de puncte de date distribuite uniform
> shapiro.test(z) # Aplicarea testului SW GoF pentru normalitate
> # Ieșire:
```

Datele testului de normalitate Shapiro–Wilk: z
W = 0,9516, valoarea p < 2,2e - 16

Observații: Valoarea p este foarte mică — mai mică de 2,2 x 10^{-16}. Prin urmare, există o concordanță slabă între valorile observate și normalitate. H_0 trebuie respins.

Note:

1. În aplicarea testului t , s-a presupus că setul de date observat a fost eșantionat dintr-o populație distribuită normal.
2. Pentru datele care au eșuat un test de normalitate, testul t poate fi folosit în continuare deoarece

(a) pentru eșantioane mici, testul t poate fi încă aplicabil deoarece distribuția statisticii t este robustă la modificări mici în ipoteza de normalitate a distribuției părinte și
(b) pentru probe mari se poate aplica teorema limită centrală, validând utilizarea testului t .

UTILIZAREA FUNCȚIEI fitdistr() PENTRU GĂSIREA VALORILOR PARAMETRILOR. R Documentația fitdistr {MASS}

Potrivirea cu maximă probabilitate a distribuțiilor univariate
Descriere:

Potrivirea cu probabilitate maximă a distribuțiilor univariate, permițând menținerea parametrilor fix dacă se dorește.

Utilizare:

fitdistr(x, densfun, start, …)

Argumente:

x Un vector numeric cu lungimea de cel puțin unul care conține numai valori *finite* .

densfun Fie un șir de caractere, fie o funcție care returnează o densitate evaluată la primul ei argument.
Distribuții „ beta ”, „ cauchy ”, „ chi-pătrat ”, „ exponențial ”, „ f ”, „ gama ”, „ geometrică ”, „ log-normal ”, „ lognormal ”, „ logistic ”, „ binom negativ ”, „ normal ”, „ Poisson ”, „ t ” și „ weibull ” sunt recunoscute, cu majusculele fiind ignorate.

start O listă numită care oferă parametrii de optimizat cu valorile inițiale . Acest lucru poate fi omis pentru unele dintre distribuțiile numite și trebuie să fie pentru altele (vezi „Detalii” de mai jos).
Parametri suplimentari, fie pentru densfun , fie pentru optim . În special, acestea pot fi folosite pentru a specifica limite prin inferioară sau superioară sau ambele. Dacă sunt incluse argumente densfun (sau funcția de densitate corespunzătoare unei specificații de șir de caractere), acestea vor fi menținute fixe.

Detalii:

> Pentru distribuțiile normale, log-normale, geometrice, exponențiale și Poisson, sunt utilizați estimatorii de maximă probabilitate în formă închisă (MLE și SE exacte), iar startul nu trebuie furnizat.

> Pentru toate celelalte distribuții, optimizarea directă a log-probabilității se realizează - folosind optim . SE-urile estimate sunt preluate din matricea de informații observate, calculată printr-o aproximare numerică. Pentru probleme unidimensionale se folosește metoda Nelder-Mead; pentru probleme multidimensionale, BFG S[20] metoda este utilizată, cu excepția cazului în care argumentele numite inferior sau superior sunt furnizate (când este utilizat L-BFGS-B) sau metoda este furnizată în mod explicit.

> Pentru distribuția numită „ t '' , densitatea este considerată ca fiind familia scară de locație cu locația m și scara s .

> Pentru următoarele distribuții denumite, se vor calcula valori inițiale rezonabile dacă startul este omis sau specificat doar parțial: „ cauchy ”, „ gamma ”, „ logistic ”, „ binom negativ ” (parametrizat prin mu și dimensiune), „ t ”, și „ weibull ”. Rețineți că aceste valori inițiale pot să nu fie suficient de bune dacă potrivirea este slabă: în special, ele nu sunt rezistente la valori aberante decât dacă distribuția ajustată este cu coadă lungă.

> Există metode print , coef , vcov și logLik pentru clasa „ fitdistr ”.

Valoare:

Un obiect din clasa „ fitdistr ”, o listă cu patru componente:

estimați estimările parametrilor
sd erorile standard estimate

[20] Algoritmul BFGS (Broyden–Fletcher–Goldfarb–Shanno) în statistică pentru probleme multidimensionale.

vcov	matricea varianță–covarianță estimată
loglik	log-probabilitatea

Următorul exemplu ilustrează utilizarea lui R pentru această aplicație.

> **Exemplul 5.19: Utilizați un** generator de numere pseudoaleatoare în R pentru a obține mostre de numere pseudoaleatoare distribuite normal și o verificare GoF pentru normalitate folosind funcția fitdistr() . Variați dimensiunile eșantionului de la 10 la 1.000.000, în multipli de 10 și verificați variațiile parametrului SD în raport cu dimensiunile eșantionului.

Soluţie:

Pot fi utilizate următoarele segmente de cod R :

```
>  # Începând cu o dimensiune a eșantionului de 1.000.000 și variază în pași de multipli de 10:
>  x <- rnorm(1000000, medie=0, sd=10)
>  fitdistr(x, „normal")
      înseamnă          sd
    -0,00503      9,99917
    (0,01000)    (0,00707)
>
>  x <- rnorm(100000, medie=0, sd=10)
>  fitdistr(x, „normal")
      înseamnă sd
    -0,0337       9,9641
    (0,0315) (0,0223)
>
>  x <- rnorm(10000, medie=0, sd=10)
>  fitdistr(x, „normal")
      înseamnă sd
    0,00829       9,90678
   (0,09907) (0,07005)
>
>  x <- rnorm(1000, medie=0, sd=10)
>  fitdistr(x, „normal")
      înseamnă sd
       0,279      9,888
      (0,313) (0,221)
>
> x <- rnorm(100, medie=0, sd=10)
> fitdistr(x, „normal")
      înseamnă sd
      0,263      8,335
     (0,834) (0,589)
>
> x <- rnorm(10, medie=0, sd=10)
> fitdistr(x, „normal")
      înseamnă sd
      -5,76      6,80
```

(2,15) (1,52)

```
> # Rezultatele sunt rezumate în Tabelul 5.12 .
> # Trasarea acestor rezultate:
> x <- c(6,80, 8,335, 9,888, 9,90678, 9,9641, 9,99917)
> y <- c(10, 100, 1000, 10000, 100000, 1000000)
> plot(y, x, ann=FALSE) # Setarea ann=FALSE pentru a inhiba adnotările
> lines(y, x, col="red") # Adăugarea unei linii roșii prin puncte
> title(main="Test GoF: Verificarea SD=10 într-o normă cu +
> SD=10", xlab="Dimensiunea eșantionului", ylab="Deviația standard")
> # Ieșire: Figura 5.7 .
```

Observatii:

- Figura 5.7 este un grafic al variației SD-urilor calculate în raport cu dimensiunile eșantionului, cu o curbă care leagă valorile calculate. Acesta arată că, pe măsură ce dimensiunile eșantionului cresc (la infinit), SD se apropie asimptotic de valoarea teoretică de 10.
- Ieșirile funcției fitdistr() au inclus SE-uri între paranteze. Acestea sunt folosite pentru a da CI pentru estimări. Aceste SE și CI 95% corespunzătoare (sd ± 1,96 × SE), au fost calculate și sunt enumerate în Tabelul 5.12 .
- Rezultatele calculate sprijină utilizarea funcției fitdistr() pentru a estima parametrii - pentru distribuțiile cunoscute și ca test GoF pentru acești parametri.

TABELUL 5.12 Utilizarea funcției fitdistr() **pentru a găsi SD într-o distribuție normală**

DIMENSIUNEA PROBA n	sd PREVIZAT DE fitdistr()	SE	CI =(sd ± 1,96 × SE)
	10,00000 (teoretic)	0	(10,00000–10,00000)
1000000	9,99917	0,00707	(9,99210–10,00624)
100000	9,9641	0,0223	(9,94180–9,98640)
10000	9,90678	0,07005	(9,83670–9,97680)
1000	9.888	0,221	(9,66700–10,10900)
100	8.335	0,589	(7,74600–8,92400)
10	6,80	1,52	(5,28000–8,32000)

Test GoF: Verificarea SD = 10 într-o normă cu SD = 10

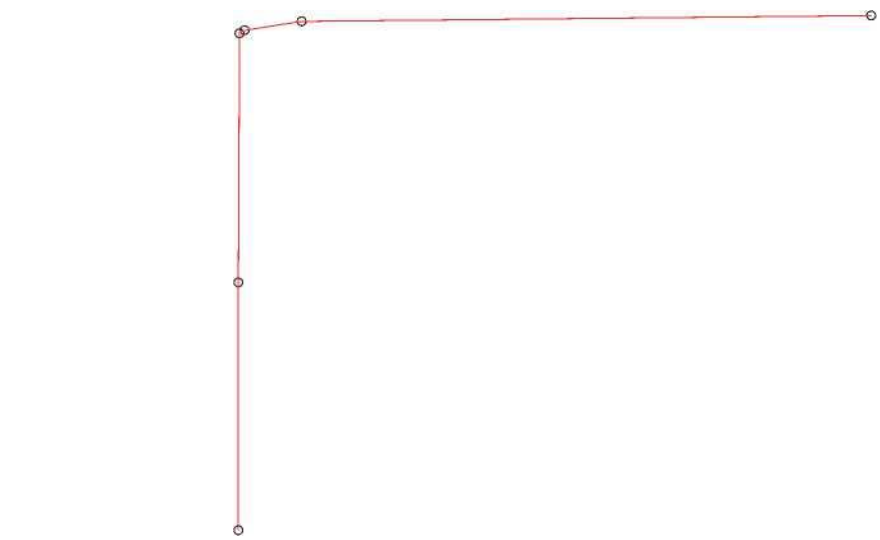


FIGURA 5.7 Utilizarea fitdistr() pentru a verifica SD într-o distribuție normală.

Mai multe informații despre aplicațiile R pentru testarea GoF pot fi găsite în pachetul CRAN actualizat gof .

Întrebări de revizuire pentru Secțiunea 5.1

1. (a) Ce este probabilitatea?
   (b) Ce este probabilitatea clasică?
2. (a) Numiți și descrieți cele trei proprietăți de bază ale probabilității.
   (b) Care este probabilitatea pentru evenimente complementare?
3. (a) Definiți următoarele, dând un exemplu pentru fiecare:
   (i) Probabilitate condiționată
   (ii) Probabilitate comună
   (b) Ce este
   (i) regula înmulțirii pentru probabilități?
   (ii) regula adunării pentru probabilități?
4. (a) Ce este probabilitatea marginală?
   (b) Dați un exemplu al acestui concept.
5. Următoarele funcții speciale R pot fi utilizate în mod convenabil în calculele probabilităților : factorial() , choose() , sample() și prod(). Dați un exemplu pentru fiecare dintre aceste funcții, așa cum sunt utilizate pentru aplicații de calcul în aplicații de probabilitate.
6. de decizie 2 × 2 care arată cele patru rezultate posibile dintr -un proces standard de testare clinică dichotomică , ce patru întrebări trebuie abordate înainte de a putea ajunge la o decizie corectă?

7. (a) Care este specificul unui test de screening sau al unui simptom de diagnostic?
   (b) Care este sensibilitatea unui test de screening sau a unui simptom de

diagnostic?

(c) Care este valoarea predictivă negativă a unui test de screening sau a unui simptom de diagnostic?

8. Descrieți pe scurt modul în care teorema lui Bayes poate fi utilizată pentru a determina sensibilitatea și specificitatea unui test de screening.
9. Folosind un exemplu, arătați că teorema lui Bayes poate fi utilizată pentru a obține valoarea predictivă pozitivă a unui test ecran.
10. Definiți și dați un exemplu pentru fiecare dintre următoarele statistici rezumative utilizate în biostatistică:
    (a) CI
    (b) Niveluri de semnificație
    (c) GoF

Exerciții pentru Secțiunea 5.1

1. Folosind R ca calculator, calculați răspunsurile la următoarele:
   (a) 7! (factorial 7)
   (b) ${}^{9}C\,5$
2. Folosind R ca calculator, calculați răspunsurile la următoarele:
   (a) Eșantionarea, fără înlocuire, a 2 seturi de 10 numere aleatorii reprezentative dintr-o populație de 2 milioane de subiecți de caz
   (b) Dintr-o populație de pacienți de 30 de persoane, lucrătorul sanitar pregătește grupuri de câte 5 pentru teste clinice ulterioare. Câte grupuri pot fi combinate, fără a avea grijă de ordinea testării în cadrul fiecărui grup?
3. (a) În ceea ce privește funcția R `fivenum()` , ce se înțelege prin rezumatul cu cinci numere din R ?
   (b) Din mulțimea primelor 100 de numere naturale, {1, 2, 3, …, *i* , …, 100}, utilizați funcția R `summary()` pentru a obține statisticile rezumative după obținerea rezumatului cu cinci numere pentru aceasta set.
4. (a) Ce se înțelege prin diagrama Q–Q a unui set de numere?
   (b) Folosind funcția R `qqnorm()` , obțineți diagrama Q–Q pentru o mulțime de 1.000 de numere distribuite normal, generate aleatoriu.
5. *Un test t cu un eșantion.*
   Aportul zilnic de calorii a 11 subiecți de caz, în kilojuli (kJ), este de 5.261, 5.674, 5.968, 6.275, 6.345, 6.587, 6.909, 7.021, 7.183, 8.251 și 8.251.
   (a) Calculați câteva biostatistici rezumative pentru acest set de date.
   (b) Aportul zilnic de energie recomandat este de 7.725 kJ. Presupunând că acest set de date face parte dintr-o distribuție normală, calculați media (*m*) a setului de date și comparați -o cu valoarea recomandată a consumului zilnic de energie.
6. *Un test Wilcoxon cu un singur eșantion.*
   (a) Pentru setul de date din Exercițiul 5, utilizați R pentru a calcula un test Wilcoxon cu un singur eșantion.
   (b) Comparați rezultatele cu cele din testul *t pentru un eșantion* . Comentează aceste două rezultate.
7. *Un exercițiu de calcul al testelor de semnificație.*

(Acest exercițiu se bazează pe o discuție a lui Dalgaard [2002, pp. 96 și urm.].)

În pachetul ISwR este un cadru de date Thuesen care constă în măsurători ale glucozei sângelui versus viteză scurtă pentru 24 de subiecți de caz. Următorul segment de cod R este utilizat pentru a obține un obiect model liniar lm pentru a reprezenta aceste două variabile , urmat de o analiză a nivelului de semnificație al corelației modelului.

completă de calcul a R este următoarea:

```
> install.packages ("ISwR")
> bibliotecă (ISwR)
```

Se atașează pachetul: „ISwR"

Următoarele obiecte sunt mascate din „pachet:supraviețuire": plămân

```
> ls("pachet:ISwR")
```

[1]	**"alkfos"**	**"ashina"**	**"bcmort"**	**„bp.obeze"**
[5]	**"cezar.pantof"**	**"cocsificabil**	**"chistfibr"**	**"eba1977"**
[9]	**"energie"**	**"ewrates"**	**„tripsină falsă"**	**„grefă.vs.gazd**
[13]	**"ritmul inimii"**	**"dracu"**	**"IgM"**	**"admisie"**
[17]	**"juul"**	**"juul2"**	**"kfm"**	**"plaman"**
[21]	**"malarie"**	**"melanom"**	**"nichel"**	**„nichel.expan**
[25]	**"filion"**	**"reacționa"**	**"celule**	**"rmr"**
[29]	**"secher"**	**"secretina"**	**"accident**	**"tb.diluat"**
[33]	**"thuesen"**	**"tlc"**	**"vitcap"**	**"vitcap2"**
[37]	**"wright"**	**"zelazo"**		

```
> date(thuesen)
> atașați (thuesen)
> Thuesen
```

	sânge.glucoză	**scurtă.viteză**
1	**15.3**	**1,76**
2	**10.8**	**1.34**
3	**8.1**	**1.27**
4	**19.5**	**1.47**
5	**7.2**	**1.27**
6	**5.3**	**1.49**
7	**9.3**	**1.31**
8	**11.1**	**1.09**
9	**7.5**	**1.18**
10	**12.2**	**1.22**
11	**6.7**	**1.25**
12	**5.2**	**1.19**
13	**19.0**	**1,95**
14	**15.1**	**1.28**
15	**6.7**	**1,52**
16	**8.6**	**N / A**

17	**4.2**	**1.12**
18	**10.3**	**1,37**
19	**12.5**	**1.19**
20	**16.1**	**1.05**
21	**13.3**	**1.32**
22	**4.9**	**1.03**
23	**8.8**	**1.12**
24	**9.5**	**1,70**

> lm(viteză.scurtă ~ glucoză în sânge)

Apel:

lm(formula = scurtă.viteză ~ sânge.glucoză)

Coeficienți:

(Intercept)	**sânge.glucoză**
1,09781	**0,02196**

> rezumat(lm(viteză.scurtă ~ glucoză.sânge))

Apel:

lm(formula = scurtă.viteză ~ sânge.glucoză) Reziduuri:

Min	**1 Î**	**Median**	**3Q**	**Max**
-0,40141	**- 0,14760**	**-0,02202**	**0,03001**	**0,43490**

Coeficienți:

	Estima	**Std. Eroare**	**valoarea t**	**Pr(> |t|)**
(Intercepta)	**1,09781**	**0,11748**	**9.345**	**6.26e-09 *****
sânge.glucoză	**0,02196**	**0,01045**	**2.101**	**0,0479 ***

Signif. coduri: 0 '*' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 " 1**

Eroare standard reziduală: 0,2167 la 21 de grade de libertate (1 observație a fost ștearsă din cauza lipsei)

R-pătrat multiplu: 0,1737, R-pătrat ajustat: 0,1343

F-statistică: 4,414 pe 1 și 21 DF, valoare p: 0,0479

> grafic (glucoză din sânge, viteză scurtă) (Figura 5.8)

> abline(lm(viteză.scurtă ~ glucoză în sânge)) (Figura 5.9)

>

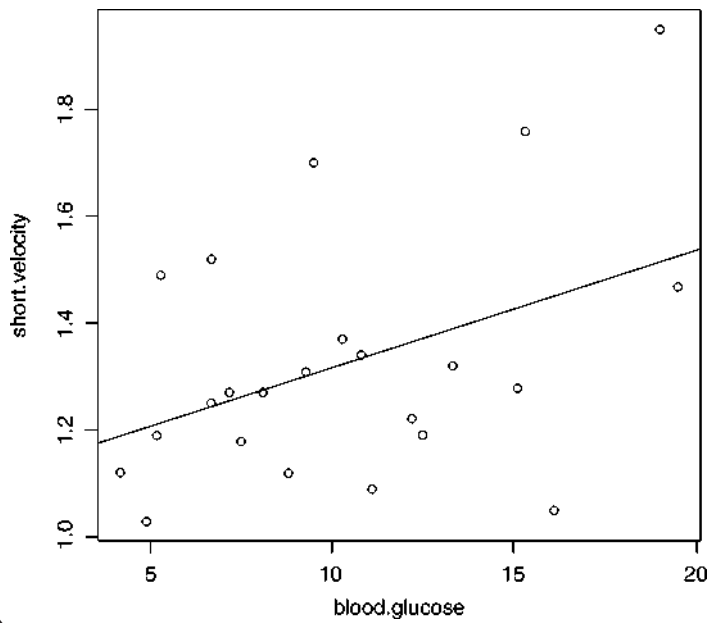
În ceea ce privește exercițiul de mai sus în calculele R :

(a) Simbolul tilde (~) din comanda > lm(viteză.scurtă ~ glucoză în sânge) poate fi citit ca „descris de". Această comandă corelează cele două variabile scurtă.viteza și sânge.glucoza . În această corelație, care componentă este
   (i) variabila *dependenta* ?
   (ii) variabila *independenta* ?

5 1015 20
sânge.glucoză

FIGURA 5.8 Diagrama de dispersie a datelor.

(b) În continuare, funcția de bază extractor summary() oferă informații cu privire la corelație. Pentru o corelare satisfăcătoare:
   (i) Media reziduurilor este, prin definiție, zero. Care este mediana reziduurilor?
   (ii) Maximul și minimul ar trebui să fie aproximativ egale în valoare absolută . Care sunt valorile absolute ale maximului și minimului rezultat?



(c) În continuare, sunt afișate coeficientul de regresie și interceptarea, însoțite de

FIGURE 5.9 Scatter plot of data with regression line.

SE, teste *t* și valori *p* . Simbolurile din dreapta tabelului sunt indicatori grafici ai nivelului de semnificație. Linia de sub tabel indică definiția acestor indicatori:

* O stea implică 0,01 < *p* < 0,05.

p calculată în acest test? Câte stele sunt?

(d) Repetați calculul, schimbând variabila dependentă cu variabila independentă. Începeți cu > lm(sânge.glucoză ~ scurtă.viteză) Comparați al doilea set de rezultate cu primul. Comentează contrastele.

8. *Un exercițiu de calcul al CI.*

În pachetul CRAN statistics este funcția R confint() , care poate fi utilizată pentru calcularea CI pentru unul sau mai mulți parametri într-un model adaptat. Există o metodă implicită și o metodă pentru obiectele care moștenesc din clasa „ lm ". Forma de utilizare pentru confint() este confint(object, parm, level = 0.95, …) , pentru care argumentele sunt:

obiect Un obiect model adaptat.

parm O specificație a căror parametri trebuie să primească CI, fie un vector de numere, fie un vector de nume. Dacă lipsesc, toți parametrii sunt luați în considerare .

nivel Nivelul de încredere necesar.

Argument(e) suplimentar(e) pentru metode.

confint() este o funcție generică. Metoda implicită presupune normalitatea asimptotică. Metoda implicită poate fi apelată direct pentru comparație cu alte metode . Pentru obiectele din clasa „ lm ", se folosesc formulele directe bazate pe valorile *t* .

(a) Calculați CI pentru obiectul model din Exercițiul 7 după următoarele segmente de cod R , pentru CI 95%:

```
> m <- lm(viteză.scurtă ~ glucoză în sânge)
> confint(m)
                              2,5 %          97,5 %
(Interceptare)        0,8534993816    1,34213037
sânge.glucoză         0,0002231077    0,04370194
```

(b) Pentru a alege 99% CI, utilizați:

```
> confint(m, nivel=0,99)
                              0,5%           99,5%
(Interceptare)        0,765183405     1,43044635
sânge.glucoză         -0.007635328    0.05156037
>
```

9. *Un exercițiu de calcul al GoF.*

În pachetul CRAN pgirmess este funcția ks.gof() , testul KS GoF pentru distribuții normale. Forma de utilizare pentru ks.gof() este ks.gof(var) , pentru care argumentul este var , un vector numeric.

Următorul segment de cod R ilustrează un calcul GoF simplu:

```
> install.packages("pgirmess")
> bibliotecă (pgirmess)
> ls("pachet:pgirmess")
[1] „CI                    classnum          cormat"
[4] „correlog              data2iarna        diag2edge"
[7] „difshannonbio         dirProj           dirSeg"
[10] „distNNeigh           distNode          distSeg"
[13] „distTot              expandpoly        friedmanmc"
[16] „gps2gpx              kruskalmc         kruskalmc.implici
[19] „kruskalmc.formula    ks.gof            pairsrp "
[22] „pavat                pclig             permcont"
[25] „Test Perm            PermTest.glm      PermTest.lm"
[28] „PermTest.lme         piankabio         piankabioboot"
[31] „plot.correlog        policirc          polycirc2"
[34] „postxt               print.clnum       print.correlog"
[37] „print.mc             print.PermTest    citeșteGDALbbox
[40] „readVista            rmls              rwhatbufCat"
[43] „rwhatbufCat2         rwhatbufNum       rwhatpoli"
[46] „Segmente             selMod            selMod.list"
[49] „selMod.lm            Shannon           Shannonbio"
[52] „shannonbioboot       tabcont2categ     thintrack"
[55] „trans2pix            trans2seg         TukeyHSD-uri"
[58] „încărcațiGPS         val4symb          valchisq"
[61] „scrie.delim          scrie GPX         scrie PRJ"
> # Să încercăm asta pe unele seturi de date distribuite normal:
> x<-rnorm(50) # Luați un set de 50 de numere generate aleatoriu
> x
 [1]  1,15482519 -0,05652142 -2,12936065  0,34484576 -1,90495545 -0,81117015
 [7]  1,32400432  0,61563685  1,09166896  0,30660486 -0,11015876 -0,92431277
[13]  1,59291375  0,04501060 -0,71512840  0,86522310  1,07444096  1,89565477
[19] -0,60299730 -0,39086782 -0,41622203 -0,37565742 -0,36663095 -0,29567745
[25]  1.44182041 -0,69753829 -0,38816751  0,65253645  1,12477245 -0,77211080
[31] -0,50808622  0,52362059  1,01775423 -0,25116459 -1,42999345  1.70912103
[37]  1,43506957 -0,71037115 -0,06506757 -1,75946874  0,56972297  1,61234680
[43] -1,63728065 -0,77956851 -0,64117693 -0,68113139 -2,03328560  0,50096356
[49] -1,53179814 -0,02499764
> ks.gof(x) # Ieșire:
        Date de testare Kolmogorov–Smirnov dintr-un eșantion: var
D = 0,0811, valoarea p = 0,8707 ipoteză alternativă: cu două fețe
```

> ks.gof(blood.glucose) **#** ***Folosind setul de date din Exercițiul 8, rezultând:***
Date de testare Kolmogorov–Smirnov dintr-un eșantion: var
D = 0,1148, valoarea p = 0,9097
ipoteză alternativă: bifață
Mesaj de avertizare:
În ks.test(var, "pnorm", mean(var), sd(var)):
legăturile nu ar trebui să fie prezente pentru testul Kolmogorov–Smirnov
> x1 <- rnorm(10000) # *Luând un set mai mare de numere* rnorm *:*
> ks.gof(x1) **#** ***Ieșire:***
Date de testare Kolmogorov–Smirnov dintr-un eșantion: var
D = 0,007, valoarea p = 0,718 ipoteză alternativă: cu două fețe
> x2 <- rnorm(1000000) **#** ***Pentru un set încă mai mare de numere rnorm:***
> ks.gof(x2)
Testul Kolmogorov–Smirnov cu o singură probă
date: var
D = 5e-04, valoarea p = 0,979
ipoteză alternativă: bifață

(a) Care sunt valorile *p* obținute prin acest test GoF pentru un set generat aleatoriu și distribuit normal: 50, 100.000 și 1.000.000?

(b) Colectați alte seturi de date și utilizați această procedură simplă pentru a testa normalitatea.

(c) Măriți dimensiunea seturilor de date și apoi repetați acest test. Cum variază progresiv valorile *p* pe măsură ce dimensiunea setului de date crește? Comentează rezultatele.

10. gof *: Software de statistică bunătate în* R *.*

O serie de pachete software statistice pentru calcule în analiza GoF sunt disponibile în mediul R cu sursă deschisă , disponibil pe site-ul web CRAN (http://cran.r-project.org)

O contribuție tipică este:

cumres : calcularea reziduurilor cumulate pentru modele liniare generalizate (GLM) în cadrul pachetului gof, care este dezvoltat ca software statistic GoF în R .

Acest software calculează măsurile GoF pentru modelele de regresie liniară lm() , inclusiv modelele de regresie logistică și Poisson, precum și modelele liniare generalizate glm() . Acestea sunt ilustrate după cum urmează:

(i) forma de utilizare a clasei „ lm ” este
cumres(model, …)

(ii) forma de utilizare a clasei „ glm ” este
cumres(model,
variabilă=c(„prevăzut”, colnames(model.matrix(model))),
data=data.frame(model.matrix(model)),
R=500, b=0, parcele=min(R,50),
semințe=rotunde(runif(1,1,1e9)),…)

în care argumentele sunt
model Obiect model (lm sau glm)
variabilă Listă de variabile după care se ordonează reziduurile
date Cadrul de date utilizat pentru a se potrivi modelului (cazuri complete)
R Numărul de mostre utilizate în simulare

b	Lățimea de bandă medie în mișcare (0 corespunde infinitului = reziduuri cumulate standard)
parcel	Numărul de realizări de salvat pentru utilizare în rutina
sămânț	
ă	Argumente adiționale aleatoare

Calculul revine ca obiect al clasei „ cumres'' .

Un exemplu de calcul este prezentat în următorul segment de cod R pentru a ilustra utilizarea acestui software GoF, cumres() , pentru a simula o funcție simplă: $f(x_1, x_2) = 10x_1 + x_2^2$, unde ambele x_1 și x_2 sunt variabile independente generate aleatoriu, distribuite normal:

```
>   instoll.pockoges(„gof'')
>   carte (gof)
```
Se încarcă pachetul „gof''…
Versiune: 0.8–1
```
>   ls("pockoge:gof")
```
[1] „cumres''
```
> sim1 <- function(n=100, f=function(x1,x2) {10+x1+x2^2},
+ sd=1, sămânță=1) {
+ dacă (!este.null(sămânță))
+ set.seed(seed)
+ x1 <- rnorm(n);
+ x2 <- rnorm(n)
+ X <- cbind(1,x1,x2)
+ y <- f(x1,x2) + rnorm(n,sd=sd)
+ d <- doto.frome(y,x1,x2)
+ return (d)
+}
>   d <- sim1(100); l <- lm(y ~ x1 + x2,d)
>   system.time(g <- cumres(l, R=100, plots=50))
```
sistemul utilizatorului a expirat
0,21 0,00 0,21
```
> g # ieșire:
```
Testul Kolmogorov–Smirnov: valoarea p = 0,32
Testul Cramer–von Mises: valoarea p = 0,36
Bazat pe 100 de realizări. Reziduuri cumulate ordonate după variabila prezisă.
--

Testul Kolmogorov–Smirnov: valoarea p = 0,51
Testul Cramer–von Mises: valoarea p = 0,26
Bazat pe 100 de realizări. Reziduuri cumulate ordonate după x1-variabilă.
--

Testul Kolmogorov–Smirnov: valoarea p = 0
Testul Cramer–von Mises: valoarea p = 0
Bazat pe 100 de realizări. Reziduuri cumulate ordonate după variabila x2.

```
> plot(g) # ieșire: Figura 5.10 .
> g1 <- cumres(l, c("y''), R=100, plots=50)
```

> g1 # *Ieșire:*
Testul Kolmogorov–Smirnov: valoarea p = 0,26
Testul Cramer–von Mises: valoarea p = 0,32
Bazat pe 100 de realizări. Reziduuri cumulate ordonate după variabila prezisă.

> plot(g1) # *Ieșire:* Figura 5.11.
> g2 <- cumres(l, c("y"), R=100, plots=50, b=0,5)
> g2 # *Ieșire:*
Testul Kolmogorov–Smirnov: valoarea p = 0,39
Testul Cramer–von Mises: valoarea p = 0,21
Bazat pe 100 de realizări. Reziduuri cumulate ordonate după variabila prezisă.

> plot(g2) # *Ieșire:* Figura 5.12. >

(a) Luați în considerare cele trei diagrame pentru g , g1 și, respectiv, g2 , prezentate în figurile 5.10 , 5.11 și 5.12. Care sunt valorile *p* corespunzătoare ale testului KS pentru aceste corelații ?
(b) Inspectați aceste trei diagrame și observați ce se întâmplă pe măsură ce valorile *p* cresc. Descrieți formele grafice ale regresiilor de corelație corespunzătoare.
(c) Dintre cele trei încercări de corelare, care oferă cea mai „cea mai bună" corelație și cea mai „proastă" corelație? De ce?

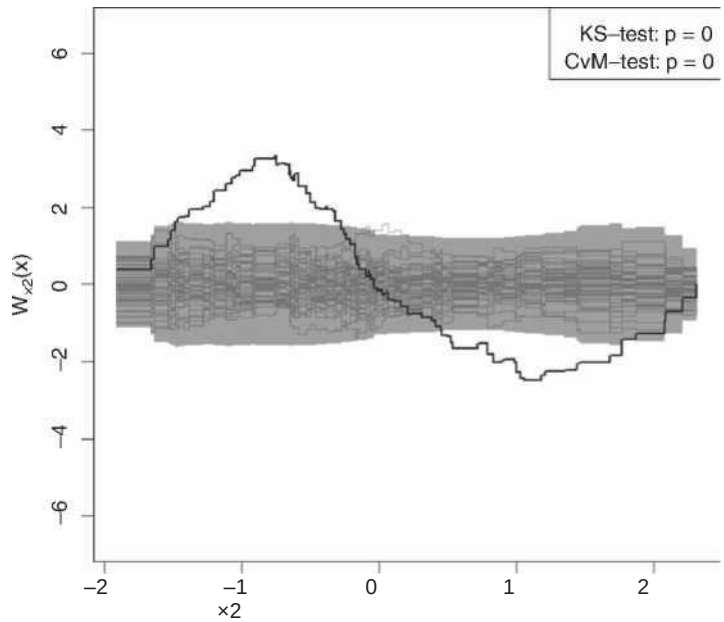


FIGURE 5.10 Plot of g.

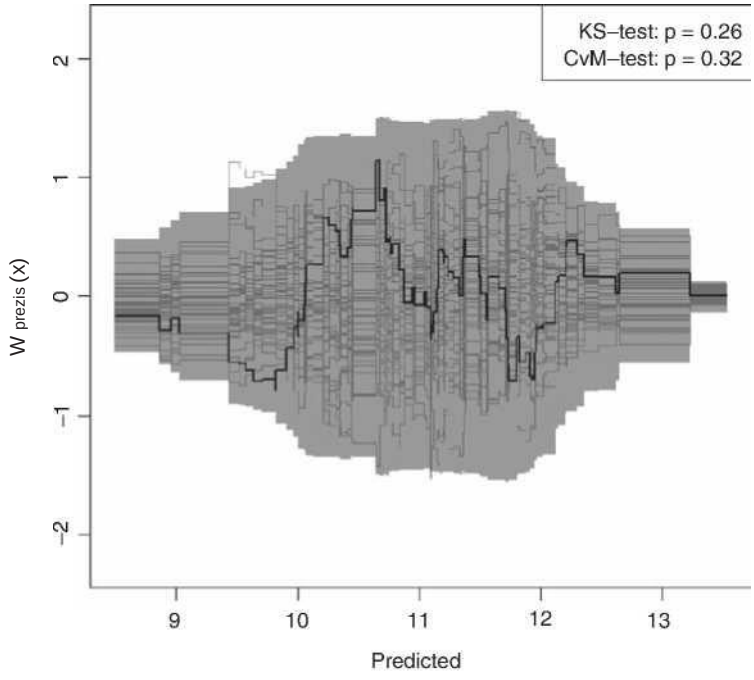


FIGURE 5.11 Plot of g1.

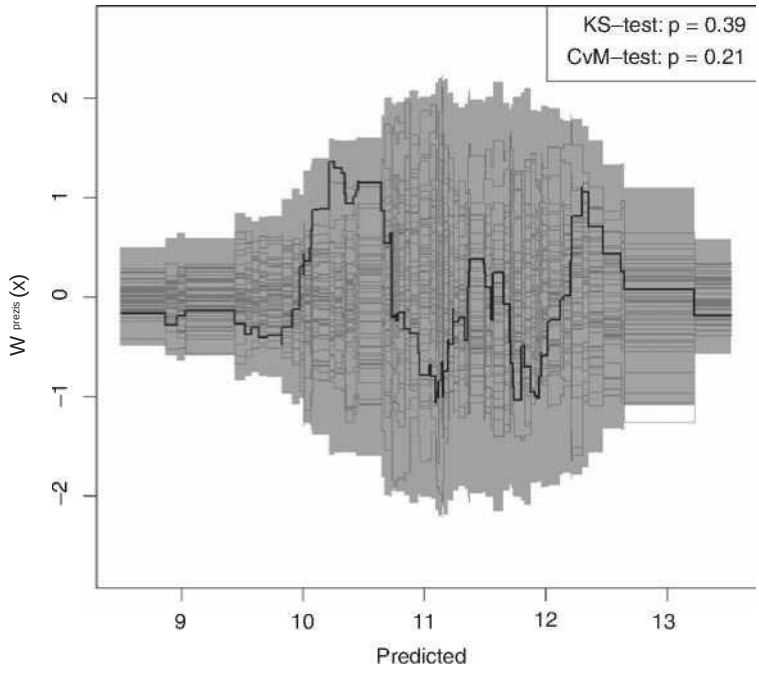


FIGURE 5.12 Plot of g2.

5.2 INFERENȚA STATISTICĂ TIPICĂ ÎN BIOSTATISTICĂ: BIOSTATISTICA BAYESIANĂ

Ce este biostatistica bayesiană?

Termenul *bayesian* se referă la Thomas Bayes (1701–1761), un matematician englez care a formulat un caz specific al teoremei care îi poartă numele.

Teoria probabilității la începutul secolului al XVIII-lea a apărut pentru a răspunde la întrebări în domeniul jocurilor de noroc și pentru a sprijini aplicațiile în noile afaceri de asigurări. Cu privire la aceasta din urmă a apărut o problemă cunoscută sub numele de întrebarea *probabilității inverse* : matematicienii vremii știau să găsească probabilitatea ca, să zicem, 5 persoane în vârstă de 50 de ani să moară într-un anumit an dintr-un eșantion de 60 dacă probabilitatea oricărui era cunoscut unul dintre ei pe moarte. Dar nu au știut cum să găsească probabilitatea ca un bărbat de 50 de ani să moară pe baza observației că 5 au murit din 60. Răspunsul a fost găsit de Bayes. Soluția sa, cunoscută sub numele de *teorema lui Bayes* , stă la baza abordării bayesiene moderne a analizei tuturor tipurilor de date.

Ancheta științifică este un proces iterativ, iar inferența bayesiană oferă un cadru logic, cantitativ pentru procesul de integrare și acumulare a informațiilor . A fost aplicat într-o multitudine de setări științifice, tehnologice și politice.

Teorema lui Bayes în teoria probabilității

Se poate vedea această teoremă seminală în termeni probabilistici.

- Forma simplă a teoremei lui Bayes
  Pentru evenimentele *A* și *B* , dacă $P(B) \neq 0$, atunci

$$P(A \cap B) = P(B \mid A)\, P(A) \,/\, P(B) \qquad (5.2\text{-}1)$$

  Din perspectiva inferenței bayesiene, probabilitatea existenței lui *B* este constantă pentru toate modelele A_n , iar posteriorul poate fi exprimat proporțional cu numărătorul:

$$P(A_n \cap B) \propto P(B \mid A_n)\, P(A_n) \qquad (5.2\text{-}2)$$

- extinsă a teoremei lui Bayes
  Pentru o partițiune a spațiului de evenimente $\{A_i\}$, care este dat sau exprimat în termeni de $P(A_i)$ și $P(B|A_i)$, se poate elimina P(B) folosind legea probabilității totale:

$$P(B) = \sum_j P(B|A_j)\, P(A_j) \Rightarrow P(A_i|B) = P(B|A_i)P(A_i)/\sum_j P(B|A_j)P(A_j) \qquad (5.2\text{-}3)$$

  Pentru cazul special al unei partiții binare:

$$P(A \mid B) = P(B|A)\, P(A)/[\, P(B \mid A)\, P(A) + P(B \mid -A)\, P(-A) \qquad (5.2\text{-}4)$$

În mod similar, pot fi găsite extensii ale teoremei lui Bayes pentru trei sau mai multe evenimente.

Metodologie bayesiană și modele de analiză de supraviețuire (timpul până la eveniment) pentru biostatistică în epidemiologie și medicină preventivă

Aplicarea teoremei lui Bayes în analiza supraviețuirii epidemiologice este o contribuție fundamentală a teoriei statistice și este considerată una dintre cele neparametrice clasice.

metode. Această abordare se bazează pe două surse: (1) credința a priori a investigatorului în caracteristica funcției de supraviețuire, atunci când este combinată cu (2) datele pentru a forma o **funcție de supraviețuire** . Aceste cunoștințe anterioare, care se pot baza pe experiențele anterioare cu comportamentul unor procese similare, precum și pe înțelegerea experților, contribuie împreună la o distribuție eficientă a funcției de supraviețuire **timp până la eveniment** .

Informațiile eșantionului sunt exprimate în termeni de funcție de probabilitate. Aceste două surse distincte de informație sunt apoi combinate de teorema lui Bayes pentru a produce distribuția a posteriori a funcției de supraviețuire, care este distribuția funcției de supraviețuire pentru datele date.

În această abordare, parametrii modelului sunt considerați ca variabile aleatorii - selectate din distribuția anterioară. Această distribuție anterioară, o distribuție multivariată asupra parametrilor, este aleasă pentru a reprezenta credința a priori a investigatorului în valorile parametrilor. Acestea reflectă cele mai bune estimări ale investigatorului cu privire la valoarea parametrilor. Varianta anterioară este o măsură a incertitudinii investigatorului în mediile anterioare. În această analiză, subiectul de interes este funcția de supraviețuire, cunoscută și sub denumirea de *funcție de hazard cumulativ* .

INFERENȚA BAYESIANĂ

Inferența bayesiană derivă *probabilitatea posterioară* ca urmare a două antecedente: (a) o *probabilitate anterioară* și (b) o funcție de probabilitate, care este derivată dintr-un model de probabilitate pentru datele de observat. Inferența bayesiană calculează apoi probabilitatea posterioară conform teoremei lui Bayes:

$$P(A \mid B) = P(B \mid A)\, P(A) / P(B) \qquad (5.2\text{-}1)$$

Când se aplică regula lui Bayes (un alt nume comun pentru teorema lui Bayes), evidența B *corespunde* datelor care nu au fost utilizate la calcularea probabilității anterioare. *A* reprezintă orice ipoteză a cărei probabilitate poate fi afectată de datele observate. (Pot exista ipoteze concurente și o decizie va fi luată pe baza probabilităților lor relative.)

Interpretarea factorilor din teorema lui Bayes este următoarea:

- *P(A* I *B),* posterior , **este** probabilitatea lui *A după ce B* este observat. Aceasta arată care sunt probabilitățile diferitelor ipoteze posibile, având în vedere dovezile observate.
- *P* (**A** *),* priorul , este probabilitatea lui *A înainte ca B* să fie observat. Acest lucru dezvăluie convingeri preconcepute despre cât de probabile sunt diferite ipoteze.
- *P (A* **I** *B)* este **probabilitatea** . Indică cât de probabil este ca cineva să observe dovezile pe care le observă de fapt, având în vedere o anumită ipoteză; cu alte cuvinte, cât de compatibile sunt dovezile cu o ipoteză dată.
- *P* (*B*) este **probabilitatea marginală** sau „dovada model". Acest factor este același pentru toate ipotezele posibile luate în considerare. (Acest lucru se poate observa prin faptul că ipoteza *A* nu apare nicăieri în simbol, spre deosebire de toți ceilalți

factori.) Aceasta înseamnă că acest factor nu intră în determinarea probabilităților relative ale diferitelor ipoteze.

Rețineți că numai factorii $P(A)$ și $P(B \mid A)$ afectează valoarea lui $P(A \mid B)$ pentru diferite valori ale lui *A.* Ambii apar la numărător și, prin urmare, probabilitatea posterioară este proporțională cu ambele. Cu alte cuvinte:

> Probabilitatea posterioară a unei ipoteze este determinată de o combinație dintre probabilitatea inerentă a unei ipoteze (precedentul) și compatibilitatea dovezilor observate cu ipoteza (probabilitatea).

Anunțat într-un mod mai concis și mai tehnic: *Posterior este proporțional cu probabilitatea timpurilor anterioare.*

Rețineți că regula lui Bayes poate fi scrisă și după cum urmează:

$$P(A \mathbf{I} B) = [P(B \mathbf{I} A) / P(B)] \times P(A) \qquad (5.2\text{-}5)$$

Factorul $[P(B \mathbf{I} A) / P(B)]$ reprezintă impactul lui *B* asupra probabilității lui *A* . Din punct de vedere logic, teorema lui Bayes are sens. Dacă dovezile nu se potrivesc cu o ipoteză, este puțin probabil să creadă ipoteza. Cu toate acestea, dacă cineva crede că o ipoteză este extrem de improbabilă a priori, este puțin probabil să o creadă, chiar dacă dovezile par să se potrivească.

De exemplu, imaginați-vă că aveți diverse ipoteze despre natura unui nou-născut. Dacă vi se prezintă dovezi sub forma unei imagini cu un băiețel cu părul negru, este probabil să credeți că copilul este într-adevăr un băiat și chiar are părul negru și este mai puțin probabil să credeți că copilul este de fapt un blond. fată cu păr, deoarece dovezile nu sunt de acord cu această din urmă ipoteză. În schimb, dacă vi se prezintă dovezi sub forma unei imagini cu un pui de câine, atunci este puțin probabil să credeți că bebelușul este de fapt un câine, deoarece credeți anterior în această ipoteză (că un om poate da naștere unui câine) este de nesuportat.

Astfel, inferența bayesiană oferă o modalitate sistematică de a combina credințele anterioare cu dovezi noi, prin aplicarea teoremei lui Bayes. Aceasta este în contrast cu inferența frecventistă, care depinde doar de dovezi în ansamblu, fără referire la credințele anterioare.

Teorema lui Bayes poate fi aplicată și în mod iterativ și repetat. După observarea unor dovezi, probabilitatea posterioară rezultată poate fi apoi tratată ca o probabilitate anterioară, iar o *nouă* probabilitate posterioară poate fi calculată din noi dovezi. Acest lucru permite ca principiile bayesiene să fie aplicate la diferite tipuri de dovezi, indiferent dacă sunt privite simultan sau în timp. Această procedură, numită *actualizare bayesiană,* este utilizată pe scară largă și eficientă din punct de vedere computațional.

ANALIZA SUPRAVIEȚIEI

În cercetările și studiile clinice, un investigator monitorizează adesea progresul subiecților de caz sub tratament dintr-un anumit moment în timp (cum ar fi atunci când se inițiază un regim de tratament medicamentos sau este întreprinsă o procedură chirurgicală critică) până la apariția și/sau reapariția un anumit eveniment (cum ar fi încetarea simptomelor critice sau moartea). De exemplu, un grup de pacienți care au avut fiecare un prim infarct miocardic (atac de cord) sunt înscriși într-o investigație de 2 ani (1 ianuarie 2011 până la 1 ianuarie 2013) pentru a evalua eficacitatea a două noi - medicamente concurente pentru prevenirea unui al doilea atac. Studiul începe atunci când primul subiect de caz, după primul atac de cord, este înscris în program și continuă până când fiecare subiect de caz experimentează unul dintre cele trei evenimente:

1. Evenimentul de interes (un atac de cord),
2. Pierderea la urmărire, din motive precum decesul dintr-o altă cauză decât un atac de cord sau încetarea participării la studiu sau
3. Încă supraviețuiește în momentul în care programul este încheiat.

Pentru fiecare subiect de caz, investigatorul înregistrează durata (în ani, luni, zile etc.) care se scurge între momentul în care subiectul cazului a intrat în studiu și în care subiectul cazului a experimentat unul dintre evenimentele de încheiere. Timpul care trece între înscriere și experimentarea unuia dintre evenimente este **timpul de supraviețuire** al subiectului . Setul de date pentru astfel de timpi de supraviețuire este **datele de supraviețuire** .

Luați în considerare următoarele informații despre patru subiecți de caz în studiul pacienților cu inimă, subiecții de caz A, B, C și D:

- Subiectul de caz A a intrat în anchetă la 1 ianuarie 2011 și avea o inimă atac din 31 decembrie 2012.
- Subiectul de caz B a intrat în studiu la 1 iulie 2011 și s-a mutat din oraș pe 31 decembrie 2011.
- Subiectul de caz C a intrat în anchetă la 1 septembrie 2011 și a murit 1 iulie 2012 dintr-o altă cauză decât un infarct.
- Subiectul de caz a intrat în program la 1 august 2011 și era încă în viață când programul de studii s-a încheiat la 31 decembrie 2012.

Prin urmare:

- Timpul de supraviețuire al subiectului A este de 24 de luni.
- Timpul de supraviețuire al subiectului B este de 6 luni. Acest lucru se numește un **timp de supraviețuire cenzurat** , deoarece evenimentul de încheiere a fost o pierdere pentru urmărire, mai degrabă decât un eveniment de interes.
- Timpul de supraviețuire al subiectului C este de 9 luni.
- Timpul de supraviețuire al subiectului D este de 17 luni. Acesta este, de asemenea, un timp de supraviețuire cenzurat.

Timpii de supraviețuire pentru subiecții cazului B și D sunt denumiți (în mod mai general) **date cenzurate.**

Timpii petrecuți în studiu de către acești patru subiecți de caz sunt reprezentați grafic în Figura 5.13 .

În studiile care compară eficacitatea a două tratamente, A și B, trei informații sunt de interes pentru fiecare subiect de caz:

1. Ce tratament, A sau B, a primit subiectul cazului?
2. Pe ce durata a fost observat subiectul cazului?
3. Subiectul cazului a experimentat evenimentul de interes în timpul studiului sau subiectul cazului a fost fie în viață la sfârșitul studiului, fie pierdut din urma urmăririi?

În studiile care nu se referă la tratamente comparative sau alte caracteristici speciale ale subiecților cazului, sunt necesari doar ultimii doi itemi. Cu aceste trei elemente de informații, se poate estima timpul median de supraviețuire al grupului care primește Tratamentul A și se poate compara cu timpul de supraviețuire median estimat al grupului

care primește Tratamentul B. Comparația celor două mediane ar oferi informații critice:

- Ce tratament poate întârzia pentru o perioadă mai lungă de timp (în medie) reapariția evenimentului de interes (în acest exemplu, apariția unui alt atac de cord)?

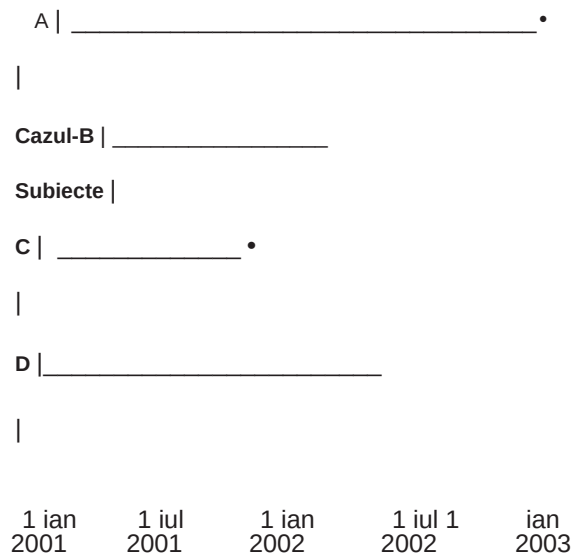


FIGURA 5.13 Patru subiecți de caz care intră într-un studiu epidemiologic în momente diferite cu timpi de supraviețuire cunoscuți (∘) și cenzurați (•).

- Care este probabilitatea estimată ca un subiect de caz să supraviețuiască pentru o anumită perioadă de timp? Clinicianul care efectuează un astfel de studiu ar putea întreba: „Care este probabilitatea ca, după primul atac de cord, un pacient care primește Tratamentul A (sau B) să supraviețuiască mai mult de doi ani?”

Metoda folosită pentru a aborda aceste tipuri de întrebări, prin utilizarea informațiilor colectate în timpul unui studiu de urmărire, se numește **analiză de supraviețuire** .

ANALIZA SUPRAVIEȚIEIULUI FOLOSIND PROCEDURA KAPLAN–MEIER. Procedura Kaplan–Meier (K–M), introdusă de Kaplan și Meier (1958), constă în multiplicarea succesivă a probabilităților individuale estimate. Este, de asemenea, cunoscută ca **metoda limită de produs** de estimare a probabilităților de supraviețuire. Procedura K–M necesită calculul proporțiilor subiecților cazului dintr-un eșantion care supraviețuiesc pentru diferite perioade de timp . Aceste proporții ale eșantionului sunt apoi utilizate ca estimări ale probabilităților de supraviețuire pe care s-ar aștepta să le observe în populația reprezentată de eșantion. În esență, acest proces estimează o funcție de supraviețuire. Distribuțiile de frecvență și probabilitate pot fi apoi construite din timpii de supraviețuire observați. Aceste distribuții observate pot arăta dovezi de conformare cu unele distribuții teoretice cunoscute.

Când forma distribuției eșantionate este necunoscută, estimarea unei funcții de supraviețuire poate fi realizată prin intermediul unei tehnici neparametrice, cum ar fi procedura K–M. Lasă:

n = numărul de subiecți de caz ai căror timpi de supraviețuire sunt disponibili

p_1 = proporția subiecților cazului care au supraviețuit cel puțin primei perioade de timp (zile, luni, ani etc.)

p_2 = proporția subiecților de caz care au supraviețuit celei de-a doua perioade de timp după ce au supraviețuit primei perioade de timp

p_3 = proporția subiecților de caz care au supraviețuit celei de-a treia perioade de timp după ce au supraviețuit celei de-a doua perioade de timp

p_k = proporția subiecților de caz care au supraviețuit celei de- *a k* -a perioade de timp după ce au supraviețuit celei de-a ($k - 1$)-a perioadă de timp

Folosind aceste proporții, putem reeticheta $\underline{p}_1$, $\underline{p}_2$, $\underline{p}_3$, …, $\underline{p}_k$ ca estimări ale probabilității ca un subiect de caz din populația reprezentată de eșantion să supraviețuiască perioadelor de timp 1, 2, 3, …, k , respectiv.

Pentru orice perioadă de timp t , $1 \le t \le k$, se poate estima probabilitatea de a supraviețui celei *de-a* -a perioade de timp, p_t , după cum urmează.

- Numărul de subiecți de caz care au supraviețuit cel puțin (t - 1) perioade de timp:

$$\underline{p}_t = \frac{\text{Numărul de subiecți de caz care supraviețuiesc, de asemenea, celei de -}a\text{-- a perioade de timp}}{\text{Numărul de subiecți ai cazului în viață la sfârșitul perioadei de timp (} t - 1)} \tag{5.2-6}$$

Atunci probabilitatea de a supraviețui până la momentul t , S (t), este estimată prin

$$\underline{S}(t) = \underline{p}_1 \times \underline{p}_2 \times \underline{p}_3 \times \cdots \times \underline{p}_t \tag{5.2-7}$$

PROCEDURA K–M CU R. R oferă două funcții care pot fi utilizate în analiza supraviețuirii:

1. Funcția survfit() , din pachetul survival
2. Funcția Surv() , tot în pachetul survival

În R , funcția survfit() creează curbe de supraviețuire fie dintr-o formulă (de exemplu, K–M), dintr-un model Cox montat anterior, fie dintr-un model de timp de eșec accelerat montat anterior. Utilizarea sa formală ia următoarea formă:

```
> survfit(formulă, …)
```

in care argumentele sunt:

Formula	Fie o formulă, fie un model montat anterior
	Alte argumente specifice metodei alese

O *curbă de supraviețuire* se bazează pe o tabelare a numărului la risc și a numărului de evenimente la fiecare moment unic al decesului. Pentru mai multe detalii, consultați documentația pentru metoda corespunzătoare: fie ?survfit.formula sau ?survfit.coxph.

De asemenea, în pachetul R survival , funcția Surv() creează un *obiect de supraviețuire* , de obicei ca răspuns într-o formulă model. Utilizarea sa ia următoarea formă:

```
> Surv(time, time2, event,
   tip=c('dreapta', 'stânga', 'interval', 'contorare', 'interval2'), origine=0)
```

in care argumentele sunt:

timp	Pentru datele cenzurate corect, acesta este timpul de urmărire. Pentru datele de interval, primul argument este ora de pornire a intervalului.

evenimen t	Indicatorul de stare – în mod normal 0 = în viață, 1 = mort. Alte opțiuni sunt TRUE/FALSE (TRUE = moarte) sau 1/2 (2 = moarte). Pentru datele cenzurate la interval, indicatorii de stare sunt 0 = cenzurat la dreapta, 1 = eveniment la momentul , 2 = cenzurat la stânga, 3 = cenzurat la interval. Deși neobișnuit, indicatorul de eveniment poate fi omis, caz în care se presupune că toți subiecții au avut un eveniment.
timp2	Ora de încheiere a intervalului numai pentru date cenzurate sau de proces de numărare. Se presupune că intervalele sunt deschise în stânga și închise în dreapta [start, end] . Pentru datele procesului de numărare, evenimentul indică dacă un eveniment a avut loc la sfârșitul intervalului.
tip	Șir de caractere care specifică tipul de cenzură. Valorile posibile sunt „ dreapta" , „ stânga" , „ numărare" , „ interval" și „ interval2" . Valoarea implicită este „ drept" sau „ contorare" , în funcție de faptul că argumentul time2 este absent sau, respectiv, prezent.
origine	Pentru datele procesului de numărare, aceasta este originea funcției de pericol. Această opțiune a fost destinată să fie utilizată împreună cu un model care conține straturi dependente de timp, astfel încât să alinieze corect subiecții atunci când trec de la un strat la altul, dar rar s-a dovedit utilă.
x	Orice obiect R.

Aplicarea procedurii K–M în mediul R este ilustrată în următoarele exemple.

> **Exemplul 5.20** Analiza supraviețuirii folosind R în radioimunoterapie cu gliom (Hastie & Tibshirani, 1990)

Un *gliom* este un tip de tumoră care începe în creier sau coloană vertebrală (cel mai frecvent în creier). Este numit astfel deoarece provine din celulele gliale.

Tratamentul pentru glioamele cerebrale depinde de localizare, tipul celulei și gradul de malignitate. Adesea, tratamentul are o abordare combinată, folosind chirurgie, radioimmunoterapia (RIT) și chimioterapie. RIT este de obicei sub formă de radiație externă.

Pentru a evalua eficacitatea clinică a RIT loco-regional, timpii de supraviețuire pentru subiecții caz dintr-un grup de control și dintr-un grup tratat (cel din urmă conținând subiecți de caz care au fost tratați cu o terapie specială) pot fi evaluați biostatistic prin reprezentarea grafică a K– M estimări ale timpilor de supraviețuire respectivi.

Folosind fișierul de date pentru gliom și aplicând procedura K–M, comparați rezultatele celor două grupuri de subiecți de caz (Histologie = Grad 3): *tratați* și *de control* . Repetați analiza numai pentru toți subiecții de sex masculin .

Soluţie:

Pot fi utilizate următoarele segmente de cod R :

```
> install.packages("survival") # Mai întâi, instalați pachetul de supraviețuire .
> library(survival) # Apoi, afișați fișierele din supraviețuire
```

Se încarcă pachetul necesar: spline

> install.packages("coin") *# Apoi, instalați pachetul* de monede .

> library(coin) *# Apoi aduceți fișierele în* coin.

Se încarcă pachetul necesar: mvtnorm
Se încarcă pachetul necesar: modeltools

Se încarcă pachetul necesar: stats4

> ls("pachet:monedă") *# Inspectarea fișierelor din* coin *pentru* gliom:

[1] „alfa”	**"alzheimer"**	**"ansari_test"**
[4] "ansari_trafo"	**"aproximativ"**	**"asat"**
[7] „asimptotic”	**"chisq_test"**	**"cmh_test"**
[10] „consal_trafo”	**"covarianta"**	**"CWD"**
[13] „dperm”	**"corect"**	**„ExactNullDistribution**
[16] „așteptare”	**"f_trafo"**	**"test_fligner"**
[19] „fligner_trafo”	**"fmaxstat_trafo"**	**"testul_friedman"**
[22] „gliom”	**"hohnloser"**	**"id_trafo"**
[25] „test_independență”	**"satisfacția**	**"testul_kruskal"**
[28] „test_lbl”	**"logrank_trafo"**	**"maxstat_test"**
[31] „maxstat_trafo”	**"testul_median"**	**"median_trafo"**
[34] „mercuryfish”	**"mh_test"**	**"neuropatie"**
[37] „test_normal”	**"normal_trafo"**	**"ocarcinom"**
[40] „of_trafo”	**"test_oneway"**	**"masina foto"**
[43] „pperm”	**"pvalue"**	**"qperm"**
[46] „rotarod”	**"testul_spearman"**	**"faza"**
[49] „statistică”	**"sprijin"**	**"surv_test"**
[52] „test_simetrie”	**"trafo"**	**"treepipit"**
[55] „varianță”	**"Wilcox_test"**	**"wilcoxsign_test"**

> attach(glioma) *# Afișarea cadrului de date* gliom

Următoarele obiecte sunt mascate din „jasa”: vârsta

> date(gliom) *# Pregătirea cadrului de date* gliom *pentru analiză* > gliom *# Privirea din interior asupra* gliomului

	nu.	vârstă	sex	histologie	grup	eveni	timp
1	**1**	**41**	**Femeie**	**Clasa 3**	**RIT**	**ADEVĂR**	**53**
2	**2**	**45**	**Femeie**	**Clasa 3**	**RIT**	**FALS**	**28**
3	**3**	**48**	**Bărbat**	**Clasa 3**	**RIT**	**FALS**	**69**
4	**4**	**54**	**Bărbat**	**Clasa 3**	**RIT**	**FALS**	**58**
5	**5**	**40**	**Femeie**	**Clasa 3**	**RIT**	**FALS**	**54**
6	**6**	**31**	**Bărbat**	**Clasa 3**	**RIT**	**ADEVĂR**	**25**
7	**7**	**53**	**Bărbat**	**Clasa 3**	**RIT**	**FALS**	**51**
8	**8**	**49**	**Bărbat**	**Clasa 3**	**RIT**	**FALS**	**61**
9	**9**	**36**	**Bărbat**	**Clasa 3**	**RIT**	**FALS**	**57**
10	**10**	**52**	**Bărbat**	**Clasa 3**	**RIT**	**FALS**	**57**

11	11	57	Masculin	Grad 3	RIT	FALS	50
12	12	55	Femeie	GBM	RIT	FALS	43
13	13	70	Bărbat	GBM	RIT	ADEVĂRAT	
14	14	39	Femeie	GBM	RIT	TRUE	14
15	15	40	Femeie	GBM	RIT	FALS	36
16	16	47	Femeie	GBM	RIT	FALS	59
17	17	58	Bărbat	GBM	RIT	ADEVĂRAT	
18	18	40	Femeie	GBM	RIT	TRUE	14
19	19	36	Barbat	GBM	RIT	TRUE	36
20	1	27	Bărbat	gradul 3	control	TRUE	34
21	2	32	Masculin	Grad 3	Control	TRUE	32
22	3	53	Femeie	Grad 3	Control	ADEVĂRAT	9
23	4	46	Bărbat	gradul 3	control	TRUE	
24	5	33	Femeie	de control	de gradul 3	FALS	50
25	6	19	Femeie	de control	de gradul 3	FALS	48
26	7	32	Femeie	GBM	Control	TRUE	8
27	8	70	Control	GBM	masculin	TRUE	8
28	9	72	Control	GBM	masculin	TRUE	11
29	10	46	Control	GBM	masculin	TRUE	12
30	11	44	Control	GBM	masculin	TRUE	15
31	12	83	Femeie	GBM	Control	TRUE	5
32	13	57	Femeie	GBM	Control	TRUE	8
33	14	71	Femeie	GBM	Control	TRUE	8
34	15	61	Control	GBM	masculin	TRUE	6
35	16	65	Control	GBM	masculin	TRUE	14
36	17	50	Control	GBM	masculin	TRUE	13
37	18	42	Femeie	GBM	Control	TRUE	25

> # *Selectarea probei necesare pentru analiză:* Grad 3 *în* histologie
> g3 <-subset(gliom, histologie == "Grad3") # *și denumindu-l* g3
> g3 # *Verificarea* g3: s-au găsit *17 subiecte* de clasa 3 *din 37 de subiecte*

	nu.	vârstă	sex	histologie	grup	timp	eveniment
1	1	41	Femeie	Grad3	RIT	TRUE	53
2	2	45	Femeie	Grad 3	RIT	FALS	28
3	3	48	Masculin	Grad 3	RIT	FALS	69
4	4	54	Masculin	Grad 3	RIT	FALS	58
5	5	40	Femeie	Grad 3	RIT	FALS	54
6	6	31	Masculin	Grad3	RIT	TRUE	25
7	7	53	Masculin	Grad 3	RIT	FALS	51
8	8	49	Masculin	Grad 3	RIT	FALS	61

9	9	36	Bărbat	Clasa 3	RIT	FALS	57
10	10	52	Bărbat	Clasa 3	RIT	FALS	57
11	11	57	Bărbat	Clasa 3	RIT	FALS	50
20	1	27	Bărbat	Clasa 3	Controla	ADEVĂRA	34
21	2	32	Bărbat	Clasa 3	Controla	ADEVĂRA	32
22	3	53	Femeie	Clasa 3	Controla	ADEVĂRA	9
23	4	46	Bărbat	Clasa 3	Controla	ADEVĂRA	19
24	5	33	Femeie	Clasa 3	Controla	FALS	50
25	6	19	Femeie	Clasa 3	Controla	FALS	48

```
>
> # Aici vine procedura KM:
> # Obțineți obiectul Surv() , apoi aplicați funcția survfit() și plot()
> plot(survfit(Surv(timp, eveniment) ~ grup, date = g3),
> main = "Gliomul de gradul III", lty = 2, col=c ("roșu", "verde"),
> xlab = „Timp de supraviețuire (luni)",
> ylab = "Valoarea probabilității",
> )
> # Emiterea curbei KM: Figura 5.14 .
```

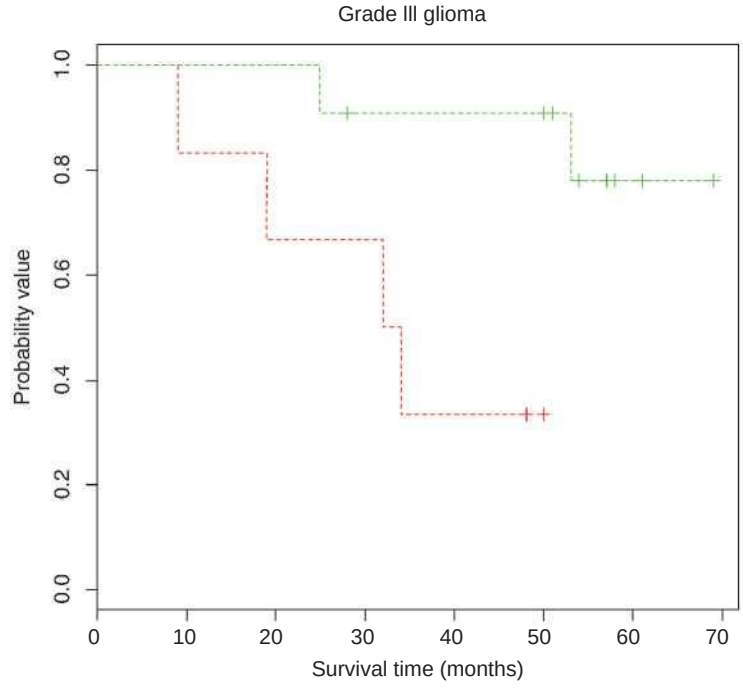


FIGURE 5.14 Bayesian survival analysis by the K–M procedure.

Using R for Glioma RIT for Grade 3 Subjects

Control Group Treated Group

Pentru a efectua aceeași analiză numai pentru toți subiecții de sex masculin , pot fi utilizate următoarele segmente de cod R :

```
>
>  g3m <- subset(gliom, sex == „Bărbat")
>  g3m # 21 Subiecți bărbați din 37 de subiecți sunt găsiți
```

	nu.	vârstă	sex	histologie	grup	evenim	timp
3	**3**	**48**	**Bărbat**	**Clasa 3**	**RIT**	**FALS**	**69**
4	**4**	**54**	**Bărbat**	**Clasa 3**	**RIT**	**FALS**	**58**
6	**6**	**31**	**Bărbat**	**Clasa 3**	**RIT**	**ADEVĂRAT**	**25**
7	**7**	**53**	**Bărbat**	**Clasa 3**	**RIT**	**FALS**	**51**
8	**8**	**49**	**Bărbat**	**Clasa 3**	**RIT**	**FALS**	**61**
9	**9**	**36**	**Bărbat**	**Clasa 3**	**RIT**	**FALS**	**57**
10	**10**	**52**	**Bărbat**	**Clasa 3**	**RIT**	**FALS**	**57**
11	**11**	**57**	**Bărbat**	**Clasa 3**	**RIT**	**FALS**	**50**
13	**13**	**70**	**Bărbat**	**GBM**	**RIT**	**ADEVĂRAT**	**20**
17	**17**	**58**	**Bărbat**	**GBM**	**RIT**	**ADEVĂRAT**	**31**
19	**19**	**36**	**Bărbat**	**GBM**	**RIT**	**ADEVĂRAT**	**36**
20	**1**	**27**	**Bărbat**	**Clasa 3**	**Controla**	**ADEVĂRAT**	**34**
21	**2**	**32**	**Bărbat**	**Clasa 3**	**Controla**	**ADEVĂRAT**	**32**
23	**4**	**46**	**Bărbat**	**Clasa 3**	**Controla**	**ADEVĂRAT**	**19**
27	**8**	**70**	**Bărbat**	**GBM**	**Controla**	**ADEVĂRAT**	**8**
28	**9**	**72**	**Bărbat**	**GBM**	**Controla**	**ADEVĂRAT**	**11**
29	**10**	**46**	**Bărbat**	**GBM**	**Controla**	**ADEVĂRAT**	**12**
30	**11**	**44**	**Bărbat**	**GBM**	**Controla**	**ADEVĂRAT**	**15**
34	**15**	**61**	**Bărbat**	**GBM**	**Controla**	**ADEVĂRAT**	**6**
35	**16**	**65**	**Bărbat**	**GBM**	**Controla**	**ADEVĂRAT**	**14**
36	**17**	**50**	**Bărbat**	**GBM**	**Controla**	**ADEVĂRAT**	**13**

```
>  plot(survfit(Surv(timp, eveniment) ~ grup, date = g3m),
+ principal = „Glioma - numai pentru bărbați", lty = 2, col=c(„roșu", „verde"),
+ xlab = „Timp de supraviețuire (luni)",
+ ylab = "Valoarea probabilității",
+ )
>  # Ieșirea curbei KM: Figura 5.15 .
```

Utilizarea R pentru Glioma RIT pentru subiecții de sex masculin

------Control Group Treated Group

Rezumat Biostatistică Decizie și Concluzie:

Curbele K–M calculate arată grafic că probabilitatea de supraviețuire pentru grupul tratat este mai mare decât pentru grupul de control. Figura 5.14 arată că pentru subiecții de gradul III, la aproximativ 48 de luni (4 ani), probabilitatea de supraviețuire pentru grupul tratat este de aproximativ 90%, comparativ cu aproximativ 30% pentru grupul de control.

Gliom - numai pentru bărbați

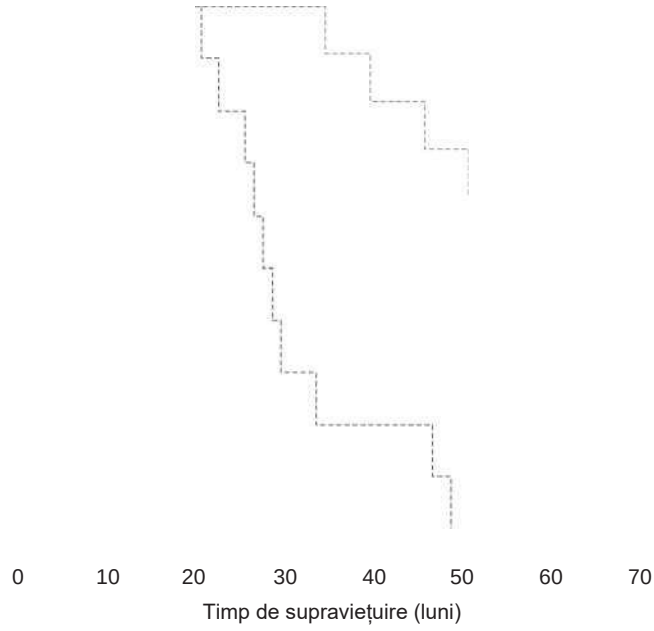


FIGURA 5.15 Analiza Bayesiană de supraviețuire prin procedura K–M.

Din Figura 5.15 , un rezultat similar este evident pentru eșantionul de bărbați. Astfel, numai pe baza analizei timpului de supraviețuire, se pare că pacienții cu gliom pot obține beneficii considerabile din tratamentul RIT; aceasta indică atât eficacitatea, cât și eficacitatea tratamentului.

Cu toate acestea, dimensiunea totală a populației este limitată și alte teste pot fi utilizate pentru a susține concluzia tentativă la care sa ajuns pe baza acestei analize K-M.

Analize suplimentare:

O serie de analize suplimentare pot fi întreprinse în R pentru a investiga în continuare (și, eventual, a sprijini) concluzia preliminară, inclusiv survfit() , survdiff() și surv-test() . Descrieri detaliate ale acestor clase de teste pot fi obținute din mediul R , folosind formatul (de exemplu, pentru survfit()): > ??survfit)

- survfit() , în pachetul survival , calculează o estimare a unei curbe de supraviețuire pentru datele cenzurate folosind fie procedura K–M, fie o altă metodă. Pentru datele privind riscurile concurente, calculează curba de incidență cumulativă.
- survdiff() , de asemenea, în pachetul survival , testează dacă există o diferență între două sau mai multe curbe de supraviețuire folosind o familie specială de teste sau pentru o singură curbă față de o alternativă cunoscută.
- surv_test() , în pachetul coin , testează egalitatea distribuțiilor de supraviețuire în două sau mai multe grupuri independente.

Următoarele sunt rezultatele respective ale acestor trei clase de teste:

1. survfit()

   > Surv (timp, eveniment)

   [1] 53 28+ 69+ 58+ 54+ 25 51+ 61+ 57+ 57+ 50+ 43+ 20 14 36+ 59+ 31 14 36
   [20] 34 32 9 19 50+ 48+ 8 8 11 12 15 5 8 8 6 14 13 25

   > survfit(Surv(timp, eveniment) ~ grup, date = g3)

Apel: survfit(formula = Surv(timp, eveniment) ~ grup, date = g3)

înregistrări		n.max	n.încep	evenim	median	0,95LCL	0,95UCL
grup=Control	6	6	6	4	33	19	N /
grup=RIT	11	11	11	2	N / A	N / A	N /

Observație: Aceste rezultate survfit() au fost utilizate în procedura K–M.

2. survdiff()

> survdiff(Surv(timp, eveniment) ~ grup, date = g3)

Apel:

survdiff(formula = Surv(timp, eveniment) ~ grup, date = g3)

	N	Observat	Așteptată	(OE)^2/E	(OE)^2/V
grup=Control	6	4	1.49	4.23	6.06
grup=RIT	11	2	4,51	1.40	6.06

Chisq= 6,1 pe 1 grad de libertate, p= 0,0138

Observație: Aceste rezultate survdiff() indică faptul că timpii de supraviețuire calculati sunt diferiți pentru cele două grupuri Control și Tratat.

3. surv_test()

> surv_test(Surv(timp, eveniment) ~ grup, date = g3)

Testul Logrank asimptotic

date: Supraveghere (timp, eveniment) pe grup (Control, RIT)

Z = 2,1711, valoarea p = 0,02992

ipoteză alternativă: cu două fețe

Observație: Aceste rezultate surv_test() indică, de asemenea, că timpul de supraviețuire calculat este diferit pentru cele două grupuri Control și Tratat. Acest test poate fi aplicat subiecților de caz cu gliom GRADE4 (GBM = *glioblastom multiform*), după cum urmează:

```
> g4 <-subset(gliom, histologie == „GBM")
> surv_test(Surv(time, event) ~ grup, date = g4,
+                 distribuție = „exact")
```

Test Logrank exact

date: Supraveghere (timp, eveniment) pe grup (Control, RIT)

Z = 3,2215, valoarea p = 0,0001588

ipoteză alternativă: cu două fețe

Observații: aceeași diferență este evidentă.

Pentru a testa dacă noul tratament este într-adevăr superior pentru ambele grupuri de tumori simultan, același test poate fi utilizat prin *stratificarea* în funcție de gradul tumorii:

```
> surv_test(Surv(timp, eveniment) ~ grup | histologie,
+                 date = gliom, distribuție = aproximativ (B = 1000000))
```

Test Logrank aproximativ

date: Surv(time, event) by

grup (Control, RIT) stratificat după histologie

Z = 3,6704, valoarea p = 7,8e - 05

ipoteză alternativă: cu două fețe

Observație: Încă o dată, rezultatele calculate sunt în consonanță cu constatările inițiale.

■ **Exemplul 5.21:** Analiza supraviețuirii folosind R: Aplicarea procedurii KM la un fișier de date

pentru leucemia mielogenă acută (AML)

Analizele de supraviețuire pot fi efectuate folosind resursele de calcul ale pachetului de supraviețuire .

(a) Utilizați funcția survfit și obțineți un rezumat al estimării supraviețuirii.
(b) Comparați și contrastați procedura K–M cu abordarea din punctul de vedere al probabilității frecventiste.

Soluţie:

Rezultatele analizelor folosind procedura K–M pot fi aplicate folosind următorul segment de cod R :

(a) Folosind funcția rezumat() :

```
> summary(leukemia.surv) # Ieșire:
```

Apel: survfit(formula = Surv(time, status) ~ x, data = aml)

x=Menținut

timp	n.risc	n.eveniment	supraviețuire	std.err	IC mai mic de 95%.	95% CI superior
9	11	1	0,909	0,0867	0,7541	1.000
13	10	1	0,818	0,1163	0,6192	1.000
18	8	1	0,716	0,1397	0,4884	1.000
23	7	1	0,614	0,1526	0,3769	0,999
31	5	1	0,491	0,1642	0,2549	0,946
34	4	1	0,368	0,1627	0,1549	0,875
48	2	1	0,184	0,1535	0,0359	0,944

x=Neintretinut

timp	n.risc	n.eveniment	supraviețuire	std.err	IC mai mic de 95%.	95% CI superior
5	12	2	0,8333	0,1076	0,6470	1.000
8	10	2	0,6667	0,1361	0,4468	0,995
12	8	1	0,5833	0,1423	0,3616	0,941
23	6	1	0,4861	0,1481	0,2675	0,883
27	5	1	0,3889	0,1470	0,1854	0,816
30	4	1	0,2917	0,1387	0,1148	0,741
33	3	1	0,1944	0,1219	0,0569	0,664
43	2	1	0,0972	0,0919	0,0153	0,620
45	1	1	0,0000	NaN	NA	NA

```
> leukemia.surv # Ieșire:
```

Apel: survfit(formula = Surv(time, status) ~ x, data = aml)

	înregistrează	n.max	n.start	evenimente	mediană	0,95LCL	0,95UCL
x=Menținut	11	11	7	31	18	NA
x=Nemenținut	12	12	11	23	8	NA

Din punctul de vedere al statisticianului frecventist , o serie de teste de semnificație pot

poate fi folosit pentru a testa ipoteza nulă, cum ar fi

$$H0 : F1 = F2 ,$$

în care F_1 şi F_2 sunt frecvenţele de apariţie pentru cazurile cu şi fără cenzură.

Funcția de supraviețuire empirică (esf): $S_n(t)$

Tratarea datelor aml *ca și cum nu ar exista cenzură,*

$$Sn(t) = \frac{NumberofObservations > t}{n} = \#t^{-1}) = \frac{}{n} \quad (5.2\text{-}8$$

În această ecuație, $S_n(t)$ este proporția de pacienți încă în remisie după t săptămâni.

Fie aml1 subsetul de date al grupului Menținut de aml . Apoi, pe o cronologie:

Calcule esf :

t	0	9	1318232831	344548	161
# (ti – t)	11	10	8765432	1	0
Sn(t)	11/11	10/11	8/11 7/11 6/11 5/11 4/11 3/11 2/11 1/11		0

Graficul acestei funcții empirice de supraviețuitor (esf), $S_n(t)$ versus t , este prezentat în Figura 5.16 .

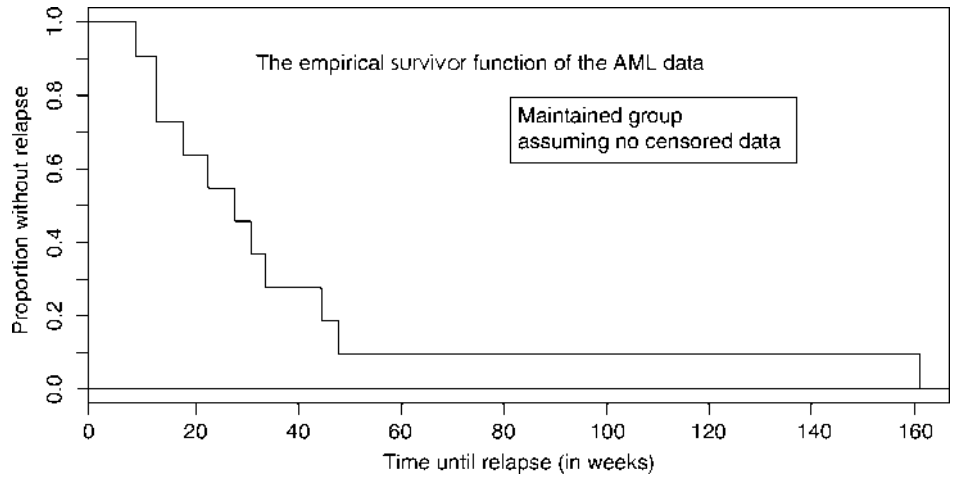


FIGURA 5.16 Funcția empirică de supraviețuitor (esf) a grupului Menținut al datelor aml1 .

Observatii:

(a) Esf este un estimator consistent al funcției de supraviețuitor adevărat **S** *(* t *)* . Distribuția exactă a *lui* $nS_n(t)$, pentru fiecare *t fix* , este binomială (n , p), unde n = numărul de observații și $p = P(T > t)$.

(b) Din teorema limită centrală, rezultă că pentru fiecare *t fix* , $S_n(t)$ este distribuit aproximativ ca Normal $\{ p , p(1 – p)/ n \}$.

(c) Esf poate fi comparat cu estimatorul produs-limită de supraviețuire, numit în mod obișnuit *estimatorul Kaplan–Meier (estimatorul K–M)* . Curba K–M este o funcție cu pas continuu drept care coboară numai la o observație necenzurată.

(d) Un grafic al curbei K–M pentru datele aml1 , împreună cu curba esf, este prezentat în Figura

5.17 . „+" de pe curba K–M reprezintă probabilitatea de supraviețuire la un moment cenzurat.

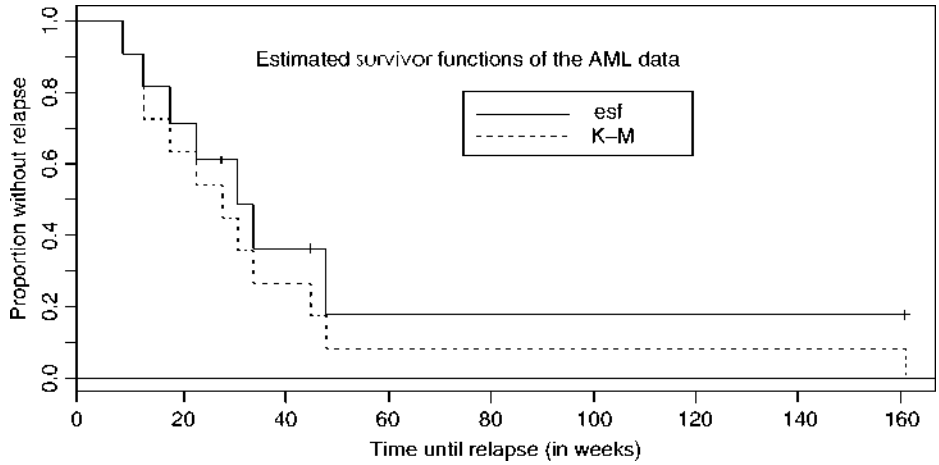


FIGURA 5.17 Estimările K–M și esf ale supraviețuirii.

(e) Observați diferența dintre cele două curbe: K–M este întotdeauna mai mare sau egal cu esf. Când nu există valori de date cenzurate, K–M se reduce la esf.

(f) Curba K–M nu ajunge până la zero, deoarece cel mai mare timp de supraviețuire (161+) este cenzurat.

Algoritmul „redistribuire la dreapta" consideră contribuția potențială a unui pacient cenzurat ca fiind redistribuită în mod egal între toți pacienții cu risc de eșec *după* timpul cenzurat.

Rezumat Biostatistică Decizie și Concluzie:

Fiecare dintre cele două afișaje grafice, figurile 5.16 și 5.17 , arată din nou că rezultatele esf le urmează îndeaproape pe cele ale procedurii K–M. Deși acesta din urmă prezintă o prezentare mai realistă , concluzia trasă din ambele abordări ar fi în general de acord.

INTRODUCERE ÎN MODELUL DE REGRESIUNE LINEARĂ MULTIPLĂ (DANIEL, 2005). În modelul de regresie liniară multiplă (ML), se *presupune* că există o relație liniară între o variabilă dependentă Y și n variabile independente: $X_1, X_2, X_3, \ldots, X_i, \ldots, X_n$.

X_i sunt cunoscute și ca *variabile explicative* sau *variabile predictoare* .

Următoarele ipoteze stau la baza analizei de regresie ML:

1. X_i sunt variabile fixe non-aleatoare (această ipoteză este în contradicție cu modelul de corelație multiplă). Această condiție înseamnă că orice inferențe extrase din datele eșantionului se aplică numai setului de valori X observate și *nu* altor colecții mai mari de X s.
2. Pentru fiecare set de valori X_i, există o subpopulare de valori Y. Pentru a construi CI și pentru a testa ipoteze, subpopulația trebuie cunoscută, cu excepția cazului în care se poate presupune că aceste subpopulații de valori Y sunt distribuite în mod normal. Această presupunere va fi făcută în primă instanță.
3. Variantele subpopulațiilor lui Y sunt toate egale.
4. Valorile Y sunt independente; adică valorile lui Y alese pentru un set de valori X nu depind de valorile lui Y alese pentru un alt set de valori X.

La generarea unei ecuații de model de regresie ML, ipotezele de mai sus pentru analiza de

regresie multiplă pot fi formulate după cum urmează:

$$y_j = b_0 + b_1 x_{1j} + b_2 x_{2j} + b_3 x_{3j} + \ldots + b_i x_{ij} + \ldots + b_n x_{nj} + e_j \quad (5,2\text{-}9)$$

unde

- y_j este o valoare tipică dintr-una dintre subpopulațiile de valori Y,
- b i sunt coeficienții de regresie ,
- x_{1j}, x_{2j}, x_{3j}, ..., x_{ij}, ..., x_{kj} sunt valori particulare ale variabilelor independente X_{1j}, X_{2j}, X_{3j}, …, X_{ij}, …, X_{nj}, respectiv, și
- e_j este o variabilă aleatoare cu medie zero și varianță s^2 , varianța comună a subpopulațiilor de valori *Y*. Pentru a seta CI pentru testarea ipotezelor despre coeficienții de regresie, presupunem că e_j sunt distribuite în mod normal și independent.

Estimările parametrilor b_0, b_1, b_2, b_3, …, b_j, …, b_k ai modelului de regresie ML specificat în ecuația (5.2-9) pot fi obținute prin *metoda celor mai mici pătrate:* suma a abaterilor pătrate ale valorilor observate ale lui *Y* de la suprafața de regresie rezultată în starea minimizată.

ANALIZA SUPRAVIEȚUIREI UTILIZAREA MODELULUI DE REGRESIUNE COX (PERICOLE PROPORȚIONALE). Tehnicile de regresie sunt disponibile atunci când măsurile dependente constau dintr-un amestec fie de date de timp până la eveniment, fie de observații de timp cenzurate . Modelul **de regresie Cox** , cunoscut și sub denumirea de *hazarduri proporționale* , este o abordare în statistica aplicată utilizată pentru a lua în considerare efectele măsurătorilor covariabile continue și discrete (variabile independente) atunci când variabila dependentă este posibil cenzurată **date** de timp până la eveniment . Modelul este, de asemenea, aplicat în mod obișnuit analizelor de supraviețuire bayesiană.

Funcția hazard, $h(ti)$, descrie probabilitatea condiționată ca un eveniment să se producă într-un moment care depășește t_i , condiționată de a supraviețui fără eveniment până la momentul t_i . Această probabilitate condiționată, cunoscută ca **rata de eșec instantanee** la momentul ti , este scrisă ca $h(ti)$. Modelul de regresie necesită să presupunem că covariatele au efectul fie de a crește, fie de a reduce riscul pentru un anumit subiect de caz, în comparație cu o anumită valoare de referință pentru funcție.

Astfel, într-un studiu clinic tipic, s-ar putea măsura *n* covariate pe fiecare dintre subiecții cazului, unde există $I = 1, 2, 3, \ldots, i, \ldots, n$ subiecți și $h_0(t_i)$ este **valoarea de bază functie de pericol** . Modelul de regresie poate fi acum scris ca

$$h(t_i) = h_0(t) \exp(b_1 z_{i1} + b_2 z_{i2} + b_3 z_{i3} + \ldots + b_i z_{ii} + \ldots + b_n z_{in}) \quad (5,2\text{-}10)$$

Coeficienții de regresie reprezintă modificările hazardului care rezultă din factorul de risc, z_{in} , care a fost măsurat. Reformarea ecuației (5.2-10) arată că coeficientul exponențiat reprezintă *raportul hazardului* sau raportul probabilităților condiționate ale unui eveniment. Aceasta este baza pentru a numi această metodă **regresia hazardelor proporționale** :

$$[h(t_i)/h_0(t)] = \exp(b_1 z_{i1} + b_2 z_{i2} + b_3 z_{i3} + \ldots + b_i z_{ii} + \ldots + b_n z_{in}) \quad (5,2\text{-}11)$$

Estimarea efectelor covariatelor, $\{b\}$, poate fi realizată cel mai bine folosind software-ul statistic , cum ar fi în mediul **R.**

MODELUL DE REGRESIUNE COX (PERICOLE PROPORȚIONALE) CU R. Atunci când este aplicat

analizei de supraviețuire, modelul de regresie Cox este similar modelelor liniare, lm , sau - modelelor liniare generalizate, glm , prin aceea că presupune liniaritate pe scara log-hazard. Modelele sunt apoi montate folosind maximizarea probabilității Cox (care *nu este* o funcție de probabilitate adevărată, deși poate fi folosită ca una). Supraviețuirea este apoi calculată ca produsul probabilităților condiționate ale timpului observat până la eveniment.

Procedura din R începe în general prin calcularea obiectelor coxph , folosind funcția coxph() , în pachetul survival . Această clasă de obiecte este returnată de clasa de funcții coxph pentru a reprezenta un model de hazarduri proporționale adaptat. Obiectele acestei clase au metode pentru funcțiile print , summary , residuals , predict și survfit .

Aplicarea modelului de regresie Cox (riscurile proporționale) în mediul R este ilustrată în exemplul următor.

■ **Exemplul 5.22:** Analiza supraviețuirii folosind R : Aplicarea modelului de regresie Cox (tehnica riscurilor proporționale) la fișierul de date despre cancerul ovarian ovarian în pachetul de supraviețuire

Acest exemplu este preluat din pachetul de supraviețuire de pe site-ul web CRAN e[21] (intitulat „ cox. zph —Test the Proportional Hazards Assumption of a Cox Regression Model Fit (coxph)," din 15 februarie 2012).

Soluţie:

Funcția R cox.zph() are următoarea formă de utilizare :

cox.zph(fit, transform="km", global=TRUE)

in care argumentele sunt:

potrivire Rezultatul potrivirii unui model de regresie Cox folosind funcția coxph() .

transform Un șir de caractere care specifică modul în care timpii de supraviețuire ar trebui să fie transformați înainte de efectuarea testului. Valorile posibile sunt „ km ", „ rank " , „ identitate " sau o funcție a unui argument.

global Întreabă dacă ar trebui efectuat un test global chi-pătrat în plus față de testele pe variabilă.

Mai mult, un obiect din clasa „ cox.zph " are următoarele componente:

tabel O matrice cu un rând pentru fiecare variabilă și, opțional, un ultim rând pentru testul global. Coloanele matricei conțin coeficientul de corelație dintre timpul de supraviețuire transformat și reziduurile Schoenfeld scalate , un chi-pătrat și valoarea *p pe două fețe* . Pentru testul global, nu există o corelație adecvată, astfel încât NA este introdusă în matrice ca deținător de loc.

x Axa temporală transformată.

y Matricea reziduurilor Schoenfeld scalate. Va fi o coloană pe fiecare variabilă și un rând pe eveniment. Etichetele rândurilor conțin timpii evenimentului inițial (pentru transformarea identității, acestea vor fi aceleași cu x).

call Secvența de apelare pentru rutină.

Calculele necesită matricea x originală a potrivirii modelului Cox. Astfel, economisește timp dacă opțiunea x=TRUE este utilizată în coxph . Această funcție este de obicei urmată atât de un

[21] Pachetul CRAN {epicalc}: http://cran.r-project.org

grafic, cât și de o imprimare a rezultatului. Graficul oferă o estimare a coeficientului beta(t) dependent de timp . *Dacă ipoteza hazardului proporțional este adevărată* , beta(t) va fi *o linie orizontală* . Imprimarea oferă un test pentru panta=0 .

Modelul de regresie Cox poate fi aplicat folosind următorul segment de cod R :

```
> install.packages(„supraviețuire")
> bibliotecă (supraviețuire)
```

Se încarcă pachetul necesar: spline

```
> ls("pachet:supraviețuire") # În căutarea cadrului de date ovarian :
[1] „aareg”              "aml"                  "attrassign"
[4] „basehaz”            "vezica urinara"       "vezica 1"
[7] „vezica urinară2”    "cancer"               "cch"
[10] „cgd”               "clogit"               "cluster"
[13] „colon”             „cox.zph”              "coxph"
[16] „coxph.control”     "coxph.detail"         "coxph.fit"
[19] „dsurvreg”          "format.Surv"          "fragilitate"
[22] „frailty.gamma”     "fragilitate.gaussian" "fragilitate.t"
[25] „inima”             „este.na.coxph.penalt  "este.na.evaluare"
[28] „is.na.Surv”        "este.evaluabil"       "este.Surv"
[31] „jasa”              "jasa1"                "rinichi"
[34] „labels.survreg”    "leucemie"             "Logan"
[37] „plămân”            "match.ratetable"      "mgus"
[40] „mgus1”             "mgus2"                "nwtco"
[43] „ovarian”           "pbc"                  "pbcseq"
[46] „pspline”           "psurvreg"             "pears"
[49] „qsurvreg”          "tabel de evaluare"    "evaluareData"
[52] „șobolani”          "cresta"               "stanford2"
[55] „strat”             „Surv”                 „survConcordanc
[58] „survdiff”          "survexp"              "survexp.mn"
```

```
   [61] „survexp.us”          "survexp.usr"              "survfit"
   [64] „survfitcoxph.fit”    "survobrien"               "survreg"
   [67] „survreg.control”     "survreg.distributions"    "survreg.fit"
   [70] „survregDtest”        "survSplit"                "tcut"
   [73] „tobin”               "tt"                       „descurcă.special
[76] „veteran”
```

> date(ovarian)
> ovarian # *Inspectarea cadrului de date* :

	futime	fustat	vârstă	rezid.ds	rx	ecog.ps
1	59	1		2	1	1
2	115	1	74,4932	2	1	1
3	156	1	66,4658	2	1	2
4	421	0	53,3644	2	2	1
5	431	1	50,3397	2	1	1
6	448	0	56,4301	1	1	2
7	464	1	56,9370	2	2	2
8	475	1	59,8548	2	2	2
9	477	0	64,1753	2	1	1
10	563	1	55,1781	1	2	2
11	638	1	56,7562	1	1	2
12	744	0	50,1096	1	2	1
13	769	0	59,6301	2	2	2
14	770	0	57,0521	2	2	1
15	803	0	39,2712	1	1	1
16	855	0	43,1233	1	1	2
17	1040	0	38,8932	2	1	2
18	1106	0	44,6000	1	1	1
19	1129	0	53,9068	1	2	1
20	1206	0	44,2055	2	2	1
21	1227	0	59,5890	1	2	2
22	268	1	74,5041	2	1	2
23	329	1	43,1370	2	1	1
24	353	1	63,2192	1	2	2
25	365	1	64,4247	2	2	1
26	377	0	58,3096	1	2	1

```
> atașează (ovarian)
> fit <- coxph(Surv(futime, fustat) ~ age + ecog.ps,
+                  date=ovarian)
> temp <- cox.zph(fit)
```

```
> print(temp) # Afișarea rezultatelor:
```

	rho	chisq	p
vârstă	**-0,243**	**0,856**	**0,355**
ecog.ps	**0,520**	**2.545**	**0,111**
GLOBAL	**N / A**	**3.195**	**0,202**

```
> plot(temp) # Trasarea curbelor.
> # Ieșire: Figura 5.18 .
>
```

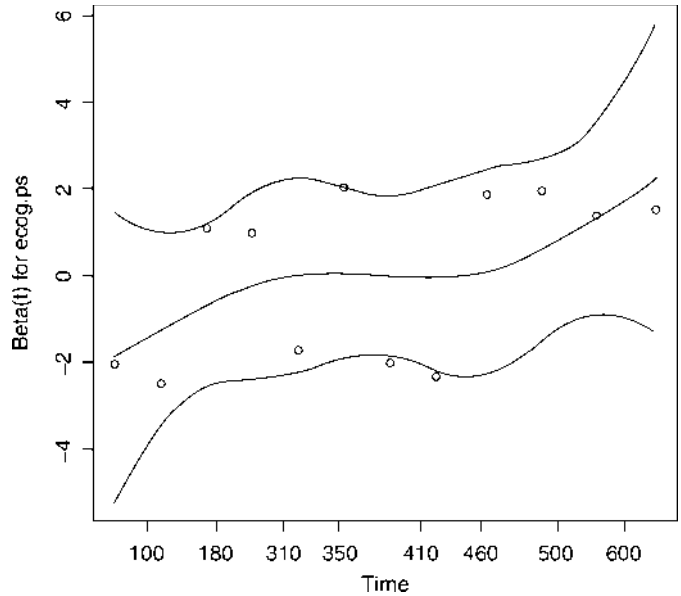


FIGURA 5.18 potrivire cox.zph() : Graficul oferă o estimare a coeficientului beta(t) dependent de timp. Dacă ipoteza hazardului proporțional este adevărată, beta(t) va fi o linie orizontală. Imprimarea oferă un test pentru panta = 0 .

Formula Bayes inversă

Formula Bayes inversă (IBF) este un instrument statistic important în distribuția teoriei și a problemelor bayesiene de date lipsă (exemplele pot fi văzute în Tan, Tian și Ng [2010] , printre altele). La începutul anului 1997, Ng a furnizat o formă de IBF în esență pentru spațiul măsurabil al produsului (PMS) și a recunoscut că IBF este potențial util în calcularea marginilor și verificarea compatibilității. Mai mult, Tian și Tan (Ng & Tong, 2010) au oferit o formă de IBF modificată în spațiul măsurabil nonprodus (NPMS) și au oferit câteva aplicații, extinzând conceptul la **formula Bayes inversă generalizată (GIBF)** .

În literatura bayesiană, se exprimă în mod tradițional distribuția posterioară în termeni de distribuție anterioară. În 1995, KW Ng a introdus **IBF** punctual pentru a-i sublinia caracterul neconvențional, prin aceea că *distribuția anterioară poate fi exprimată în termeni de distribuție posterioară* .

În notația bayesiană standard, se folosește $\pi(q)$ pentru a desemna funcția anterioară de densitate de probabilitate (pdf) a parametrului q cu suport $S (\Theta)$; $L (y | \mu)$ pentru a desemna funcția de probabilitate (adică, pdf-ul datelor, dat fiind parametrul) cu suport $S (Y | q)$; $p (q | y)$ pentru a desemna pdf- ul posterior cu suport $S(\Theta \mathbf{I} Y)$ al parametrului dat; și $f (y)$ pentru a desemna pdf necondiționat pentru datele cu suport $S (Y)$. Atât q și y pot fi vectori.

Rețineți că, în general, proiecția lui $S (Y|q)$ în $S (Y)$ este o submulțime:

$$S (Y | q) \subset S (Y) \tag{5,2-12A}$$

si egalitatea

$$S (Y | q) = S (Y) \tag{5.2-12B}$$

poate ține pentru unele q .

În ceea ce privește integrala sau probabilitatea, aceasta din urmă este în esență aceeași ca atunci când complementul proiecției lui $S (Y|q)$ în $S (Y)$ este o mulțime de măsură zero. Dacă suportul de îmbinare $S (q , Y)$ este egal cu spațiul produs $S (\Theta) \times S (Y)$, atunci:

$$S (Y|q) = S (Y) \tag{5,2-13}$$

pentru toate q și invers. O relație similară este valabilă între $S (Q \mathbf{I} y)$ și $S (Q)$. Din identitatea pdf comună, $L (y|q) p (q) = p (q | y) f (y)$, formula Bayes

$$p(q|y) = p(q)L(y|q)/ \int_{S(q|y)} p(q)L(Y|q)dq \tag{5.2-14}$$

rezultă printr-o substituție a lui $f (y)$, care este exprimată ca integrală a articulației pdf în raport cu q peste $S (Q \mathbf{I} y)$. Se poate rescrie această identitate pdf comună ca $p (q) L (y \mathbf{I} q) / p (q | y) = f (y)$, unde (q, y) este în suportul comun $S (\Theta , Y)$. Acum, pentru orice q fix , se pot integra ambele părți ale identității pdf comune re-exprimate în raport cu y peste $S (Y | q)$ și se pot obține pdf-ul anterior la q :

$$p (q) = \int_{S(Y|q)} f (y) dy \left\{ \int_{S(Y|q)} \left[L (y | q) p (q | y) \right]^{-1} \right\} \tag{5.2-15A}$$

$$\leq \left\{ \int_{q} \left[L (y | q) p (q | y) \right]^{-1} \right\} \tag{5.2-15B}$$

unde egalitatea este valabilă dacă și numai dacă $S (Y|q) = S (Y)$, sau complementul proiecției lui $S (Y | q)$ în $S (Y)$ este o mulțime de măsură zero.

În special, sub așa-numita ipoteză de pozitivitate, unde

$$S (Q , Y) = S (Q) \times S (Y) \tag{5.2-16}$$

unul are

$$p (q) = \left\{ \int_{q} \left[L (y | q)/ p (q / y) \right]^{-1} dy \right\} \tag{5.2-17}$$

Forma explicită a ecuației (5.2-17) nu a fost găsită în literatura generală de statistică bayesiană clasică. Acest lucru se poate datora tradiției din literatura bayesiană de a exprima distribuția posterioară în termeni de distribuție anterioară. Ng (2010)

numită Ecuația (5.2-17) **IBF** (punctual) pentru a sublinia caracterul său neconvențional, prin aceea că *distribuția anterioară este exprimată în termeni de distribuție posterioară* . De fapt, este media armonică a lui $p(q \mid y)$ în raport cu $L(y \mid q)$.

DERIVAȚII A TREI FORME ALE FORMULEI INVERSE BAYES (TAN, TIAN și NG, 2010)

Din formula Bayes, se pot deriva trei IBF care implică integrarea în suportul unei variabile aleatoare. Pentru a face acest lucru, este necesar să luăm în considerare două noțiuni: **spațiul măsurabil produs (PMS)** și **spațiul măsurabil nonprodus (NPMS)** .

Fie două variabile aleatoare (sau vectori) (X , Y), luând valori în spațiile ($\boldsymbol{X}$, $\boldsymbol{Y}$), respectiv, absolut continue față de o anumită măsură m pe suportul *de îmbinare* :

$$\boldsymbol{S} = \boldsymbol{S}_{(X,\ Y)} = \{(x, y): f_{X,\ Y)}(x, y) > 0, (x, y) \in (\boldsymbol{X}, \boldsymbol{Y})\} \tag{5,2-18}$$

unde $f_{(X,Y)}(x, y)$ denotă funcția de densitate de probabilitate comună (pdf) a lui (X , Y).

Acum, notează pdf-urile marginale și condiționale ale lui X și Y prin $f_X(x)$, $f_Y(y)$, $f_{(X|Y)}(x|y)$ și $f_{(Y|X)}(y|x)$, respectiv. Lasă

$$\boldsymbol{S}_X = \{x : f_X(x) > 0, x \in \boldsymbol{X}\} \tag{5,2-19}$$

și

$$\boldsymbol{S}_Y = \{y : f_Y(y) > 0, y \in \boldsymbol{Y}\} \tag{5.2-20}$$

notează suporturile lui X și respectiv Y.

Dacă

$$\boldsymbol{S}_{(X,\ Y)} = \boldsymbol{S}_X \times \boldsymbol{S}_Y \tag{5,2-21}$$

atunci măsura m este *produsul măsurabil* și m poate fi scris ca $m_X \times m_Y$; în caz contrar, nu este *măsurabil în produs* .

Ipoteza continuă absolută permite luarea în considerare a unei variabile continue (a cărei densitate este măsurabilă Lebesgue) cu o variabilă discretă (funcția sa de masă de probabilitate dă naștere unei măsuri de numărare). Pentru a aplica acest concept, se pot desemna suporturile condiționale ale lui $X|(Y = y)$ și $Y|(X = x)$ după cum urmează:

$$\boldsymbol{S}_{(X\ i\ Y)}(Y) = \{x : f : {}_{X\ i\ Y)}(x^{1}Y) > 0, x \in \boldsymbol{X}\} \ \forall\ Y \in S_Y \tag{5,2-22}$$

$$\boldsymbol{S}_{(Y\ i\ x)}(x) = \{Y : f_{(Y\ i\ x)}(Y \mid x) > 0, Y \in \boldsymbol{Y}\} \ \forall\ x \in S_X \tag{5,2-23}$$

În practică, de obicei, cineva are

$$\boldsymbol{S}_{(Y\ i\ x)}(x) \subseteq S_Y \mathrm{V}\, x \in S_x \tag{5,2-24}$$

și

$$\boldsymbol{S}_{(X\ i\ Y)}(x) \subset S_X \mathrm{V}\, Y \in S_Y \tag{5,2-25}$$

astfel încât pdf-ul comun devine

$$f;_{x\ i\ Y)}(x \mid Y) f_y(Y) = f;_{Y\ i\ x)}(Y \mid x) f_x(x), \forall (x, Y) \in \boldsymbol{S}_{(x,\ y}$$

) (5,2-26)

In PMSs, one has

$$S_{(Y|X)}(x) = S_Y, \forall x \in S_X \tag{5.2-27}$$

and

$$S_{(X|Y)}(x) = S_X \forall y \in S_Y \tag{5.2-28}$$

Prin urmare, din ecuația (5.2-26) , se obține prin împărțire:

$$f_Y(y) = \left[f_{(Y|X)}(y|x) / f_{(X|Y)}(x|y) \right] f_X(x), \forall x \in S_X \text{ și } \forall y \in S_Y \quad (5,2\text{-}29A)$$

Integrând identitatea (5.2-29A) față de *y*, pe suport S_Y, se obține:

$$\int f_Y(y)\,dy = \int \left\{ \left[f_{(Y|X)}(y|x) / f_{(X|Y)}(x|y) \right] f_X(x) \right\}$$

$$1 = \int \left\{ \left[f_{(Y|X)}(y|x) / f_{(X|Y)}(x|y) \right] f_X(x) \right\} dy$$

$$= f_X(x) \left\{ \int \left[f_{(Y|X)}(y|x) / f_{(X|Y)}(x|y) \right] dy \right\}$$

Deoarece *x* și *y* sunt independente, se obține

$$f_X(x) = \left\{ \int_S \left[f_{(Y|X)}(y|x) / f_{(Y|X)}(x|y) \right] dy \right\}^{-1}, \forall x \in S_x \tag{5.2-30A}$$

care este IBF punctual pentru X (x).

Forma duală a ecuației (5.2-29A) este

$$f_X(x) = \left[f_{(X|Y)}(x|y) / f_{(Y|X)}(y|x) \right] f_Y(y), \forall x \in S_X \text{ și } \forall y \in S_Y \tag{5.2-29B}$$

întrucât forma duală a ecuației (5.2-30A) este

$$f_Y(y) = \left\{ \int_S \left[f_{(X|Y)}(x|y) / f_{(Y|X)}(y|x) \right] dx \right\}^{-1}, \forall y \in S_Y \tag{5.2-30B}$$

La înlocuirea $f_Y(y)$ din ecuația (5.2-30B) în ecuația (5.2-29B), se obține

$$f_X(x) = \left\{ \int \left[f_{(X|Y)}(x|y) / f_{(Y|X)}(y|x) \right]^{-1} dx \right\} \left[f_{(X|Y)}(x|y) / f_{(Y|X)}(y|x) \right], \forall (x, y) \in \begin{matrix} S_{(X,Y)} \\ S_x \end{matrix}$$

$$= \left\{ \int_S \left[f_{(X|Y)}(x|y_0) / f_{(Y|X)}(y_0|x) \right]^{-1} dx \right\} \left[f_{(X|Y)}(x|y_0) / f_{(Y|X)}(y_0|x) \right]$$

$$\forall x \in S_X \text{ și } \forall \text{ fixat arbitrar } y_0 \in S_Y \tag{5.2-31A}$$

care este **IBF funcțional** pentru X (x).

În cele din urmă, la omiterea constantei de normalizare din ecuația (5.2-31A) , se obține

$f_X(x) \propto \left[\lfloor f_{(X|Y)}(x \mid y_0) / f_{(Y|X)}(y_0 \mid x) \right] \rfloor$, $\forall\ x \in S_X$ și $\forall$ fixat arbitrar $y_0 \in S_Y$ (5.2-32A) care este **IBF de eșantionare** pentru $X(x)$.

În mod clar, formele duale corespunzătoare ale ecuațiilor (5.2-30A), (5.2-31A) și (5.2-32A) pot fi formulate cu ușurință, pentru $f_Y(y)$, prin interschimbarea $X(x)$ cu $Y(y)$, ca IBF-uri corespunzătoare pentru $Y(y)$:

$$f_Y(y) = \left\{ \lfloor \int_S \left[\lfloor f_{(X|Y)}(x \mid y) f_{(Y|X)}(y \mid x) \right] \rfloor dx \right\}^{-1} \rfloor,\ \forall\ y \in S_Y \quad (5.2\text{-}30B)$$

$$f_Y(y) = \left\{ \lfloor \int_S \left[\lfloor f_{(Y|X)}(y \mid x_0) / f_{(X|Y)}(x_0 \mid y) \right] \rfloor dy \right\}^{-1} \rfloor \left[\lfloor f_{(Y|X)}(y \mid x_0) / f_{(X|Y)}(x_0 \mid y) \right] \rfloor,$$

$$\forall\ y \in \boldsymbol{S}_Y \text{ și } \forall \text{ fix arbitrar } x_0 \in \boldsymbol{S}_X \quad (5.2\text{-}31B)$$

și

$f_Y(y) \propto \left[\lfloor f_{(Y|X)}(y \mid x_0) / f_{(X|Y)}(x_0 \mid y) \right] \rfloor$, $\forall\ y \in S_Y$ și $\forall$ fix arbitrar $x_0 \in S_X$ (5,2-32B)

IBF punctual corespunzătoare pentru $Y(y)$, **IBF funcțional** pentru $Y(y)$ și respectiv **IBF de eșantionare** pentru $Y(y)$.

Următorul exemplu ilustrează aplicarea IBF pentru obținerea probabilităților anterioare din probabilitățile posterioare.

■ **Exemplul 5.23:** O aplicare a formulei Bayes inversă (Ng & Tong, 2010)

Acest exemplu arată o aplicare directă a **IBF punctual** pentru $Y(y)$, folosind ecuația (5.2-30B) pentru a obține probabilitatea anterioară a *lui* $f_Y(y)$ din informațiile disponibile în probabilitățile posterioare ale sistemului.

Luați în considerare următoarele densități de probabilitate condiționată:

$$f_{XY}(X\ Y = 1/2\sqrt{(1 - y^2)}$$

$$-\sqrt{(1 - y^2)} < x < \sqrt{(1 - y^2)}$$

$$-1 < y < 1$$

şi

$$f_{YX}(y \mid x) = 1/2\sqrt{(1 - x^2)}$$

$$-\sqrt{(1 - x^2)} < y < \sqrt{(1 - x^2)}$$

$$-1 < x < 1$$

Suportul S_{XY} este *interiorul* discului unității: cu centru (0, 0) și rază 1 unitate. $f_{X|Y}(x \mid y) / f_{Y|X}(y \mid$

x) este definit doar pe S_{XY} , de-a lungul circumferinței discului unitar.

Pentru a calcula $f_Y(y)$, considerăm o funcție de extensie pozitivă a *lui* $f_{X|Y}(x \mid y)/ f_{Y|X}(y \mid x)$ pe spațiul $(-1, 1) \times (-1, 1)$. Set

$$u(\chi) = \sqrt{(1 - \chi^2)}, \text{ pentru } 1 < x < 1; \text{ și } v(y) = 1/\sqrt{(1 - y^2)}, \text{ pentru } 1 < y < 1$$

Atunci, $r(x, y) = u(x)\, v(y)$ poate fi considerat ca reprezentând $f_{X|Y}(x \mid y) / f_{Y|X}(y \mid x)$ pe spațiul $(-1, 1) \times (-1, 1)$. Mai mult, aplicând IBF punctual folosind ecuația (5.2-30B), obținem

$$f_Y(y) = \left(\int_{S_X} \Gamma f; {}_{X I Y)}({}_{x_i} y) / f_{(Y i x)}(y \text{ eu } x \, Я^{dx} \right)^{-1} \text{f}, \forall\ y \in S_Y \qquad (5,2\text{-}30B)$$

$$f_Y(y) = \int {}_1 U(x)\, v(y)^{-1} dx$$

$$\int {}_{1\,1} \sqrt{}(1 - x^2) / 7(1 - y^2)\, dx$$

$$= 2_y \ / \ (1 - y^2) / p\,, \text{ forali } -1 < y < 1, \text{ după caz}$$

> **Exemplul 5.24:** O aplicare a simulării distribuției posterioare folosind R

În cadrul pachetelor CRAN disponibile în prezent, există pachete de aplicații care utilizează simularea posterioară pentru a obține inferența bayesiană. De exemplu, funcția MCMCregress() , din pachetul MCMCpack , folosește Markov Chain Monte Carlo pentru regresia liniară Gaussiană.

Această funcție MCMCregress() generează un eșantion din distribuția posterioară a unui model de regresie liniară cu erori Gaussiene utilizând eșantionarea Gibbs (cu o variantă Gaussian anterioară multivariată pe vectorul beta și o gama inversă înainte de varianța erorii condiționate). Utilizatorul trebuie să furnizeze date și priorități, iar un eșantion din distribuția posterioară este returnat ca obiect mcmc , care poate fi ulterior analizat cu funcțiile furnizate în pachetul coda .

Contextul teoretic al aplicării funcției

MCMCregress()

MCMCregress() simulează din distribuția posterioară folosind eșantionarea Gibbs standard (o retragere Normală multivariată pentru beta și o retragere Gamma inversă pentru varianța erorii condiționate). Simularea propriu-zisă este efectuată în cod C++ compilat pentru a maximiza eficiența; consultați documentația coda pentru o listă cuprinzătoare de funcții care pot fi utilizate pentru a analiza eșantionul posterior.

Modelul ia următoarea formă:

$$y_i = x'_I b + e_i \qquad (5,2\text{-}33)$$

unde erorile sunt presupuse a fi gaussiene, după cum urmează:

$$e_i \sim \mathrm{N}\ (0, s^2) \qquad (5,2\text{-}34)$$

Presupunând antecedente standard, semiconjugate:

$$B \sim \mathrm{N}\ (b_0, B_{0^{-1}}) \qquad (5,2\text{-}35)$$

şi

$$s^2 \sim \textit{Gamma}\ (c_0/2, d_0/2) \qquad (5,2\text{-}36)$$

unde b și s^2 sunt presupuse a priori independente.

Rețineți că sunt permise numai valorile de pornire pentru b , deoarece simularea se face

utilizând eșantionarea Gibbs cu varianța de eroare condiționată ca prim bloc din eșantioner.

Valoarea funcției MCMCregress() este un obiect mcmc care conține eșantionul posterior. Acest obiect poate fi rezumat prin funcțiile furnizate de pachetul coda .

Codul R pentru acest exemplu este

```
linie <- lista(X = c(-2,-1,0,1,2), Y = c(1,3,3,3,5))
posterior <- MCMCregress(Y~X, date=line, verbose=1000)
complot (posterior)
raftery.diag(posterior)
rezumat (posterior)
```

Acest exemplu este rulat folosind următorul segment de cod R :

```
> install.packages(„MCMCpack")
> bibliotecă (MCMCpack)
Se încarcă pachetul necesar: coda
Se încarcă pachetul necesar: zăbrele
Se încarcă pachetul necesar: MASS
##
> # Pachet Monte Carlo cu lanț Markov (MCMCpack)
> # Copyright (C) 2003-2012 Andrew D. Martin, Kevin M. Quinn și Jong Hee Park
##
> # Sprijin oferit de Fundația Națională pentru Știință din SUA
> # (Grants SES-0350646 și SES-0350613)
##
> ls("pachet:MCMCpack") # Se notează funcția MCMCregress()
 [1] „BayesFactor"          "choicevar"              "ddirichlet"
 [4] „dinvgamma"            "dorință"                "dnoncenhypergeom"
 [7] „dtomogplot"           "dorinta"                „HMMpanelFE"
[10] „HMMpanelRE"           „face.listă de pauză"    "MCbinomialbeta"
[13] „MCMCbinaryChange"     "MCMCdynamicEI"          „MCMCdynamicIRT1d"
[16] „MCMCfactanal"         "MCMChierEI"             „MCMChlogit"
[19] „MCMChpoisson"         „MCMChregress"           „MCMCirt1d"
[22] „MCMCirtHier1d"        „MCMCirtKd"              „MCMCirtKdHet"
[25] „MCMCirtKdRob"         „MCMClogit"              „MCMCmetrop1R"
[28] „MCMCmixfactanal"      „MCMCmnl"                „MCMCoprobit"
[31] „MCMCoprobitChange"    „MCMCordfactanal"        "MCMCpoisson"
[34] „MCMCpoissonChange"    „MCMCprobit"             „MCMCprobitChange"
[37] „MCMCquantreg"         " MCMCregress "          „MCMCSVDreg"
[40] „MCMCtobit"            „MCmultinomdirichlet"    "MCnormalnormal"
[43] „MCpoissongamma"       "mptable"                "plotChangepoint"
[46] „plotState"            „PostProbMod"            "procruste"
[49] „rdirichlet"           "citește. coasă"         "rinvgamma"
[52] „riwish"               "rnoncenhypergeom"       "rwish"
[55] „SSVSquantreg"         „testpanelGroupBreak"    „testpanelSubjectBreak
[58] „tomogplot"            "topmodele"              "vech"
[61] „scrie.Coasă"          "xpnd"
> bibliotecă (zăbrele)
> bibliotecă (coda)
> bibliotecă (MASS)
```

```
> linie <- lista(X = c(-2,-1,0,1,2), Y = c(1,3,3,3,5))
> posterior <- MCMCregress(Y~X, date=line, verbose=1000)
> # ieșire:
```

MCMCregress iterația 1 din 11000
beta =
3.11989
0,80906
sigma2 = 0,27278

Iterația MCMCregress 1001 din 11000
beta =
2,87038
0,80699
sigma2 = 0,38403

Iterația MCMCregress 2001 din 11000
beta =
3,75130
0,44314
sigma2 = 0,77820

MCMCregress iterația 3001 din 11000
beta =
3,72461
0,97119
sigma2 = 1,33523

Iterația MCMCregress 4001 din 11000
beta =
2,79921
0,58334
sigma2 = 0,80892

Iterația MCMCregress 5001 din 11000
beta =
3,06049
0,99932
sigma2 = 0,35595

Iterația MCMCregress 6001 din 11000
beta =
3,09242
0,68169
sigma2 = 1,37335

Iterația MCMCregress 7001 din 11000
beta =
3,21538
1,35860
sigma2 = 5,18619

Iterația MCMCregress 8001 din 11000

beta =
3,13608
0,81812
sigma2 = 1,46922

Iterația MCMCregress 9001 din 11000
beta =
3,40301
1,43191
sigma2 = 2,76159

Iterația MCMCregress 10001 din 11000
beta =
4,78574
2,30508
sigma2 = 3,57808

> plot(posterior) *# Ieșirea mai multor diagrame, rezumate în* Figura 5.19 .
> raftery.diag(posterior) *# Ieșire:*

Quantila (q) = 0,025
Precizie (r) = +/-0,005
Probabilitate (s) = 0,95

	Ardere (M)	Total (N)	Limită inferioară (Nmin)	Factorul de dependență
(Intercepta)	**3**	**4374**	**3746**	**1.17**
X	**3**	**4374**	**3746**	**1.17**
sigma2	**2**	**3865**	**3746**	**1.03**

> rezumat(posterior)
Iterații = 1001:11000
Interval de subțiere = 1
Număr de lanțuri = 1
Dimensiunea eșantionului per lanț = 10000

1. Media empirică și abaterea standard pentru fiecare variabilă, plus standard eroarea mediei:

				Seria temporală
	Medie	**SD**	**SE naiv**	**SE**
(Intercepta	**3.013**	**0,5459**	**0,005459**	**0,005948**
X	**0,807**	**0,3873**	**0,003873**	**0,004647**
sigma2	**1.510**	**5,0969**	**0,050969**	**0,084921**

2. Cuantile pentru fiecare variabilă:

	2,5%	25%	50%	75%	97,5%
(Intercepta)	**2,04394**	**2,7552**	**3,0043**	**3,2483**	**4.083**
X	**0,05681**	**0,6287**	**0,8047**	**0,9769**	**1.562**
sigma2	**0,17091**	**0,3876**	**0,6837**	**1,3395**	**7.155**

>

```
> linie <- lista(X = c(-2,-1,0,1,2), Y = c(1,3,3,3,5))
> linia # ieșire:
$X
[1] -2 -1 0 1 2

$Y
> 1] 1 3 3 3 5
> posterior <- MCMCregress(Y~X, date=line, verbose=1000)
>  complot (posterior)
>  # ieșire: Figura 5.19 .
```

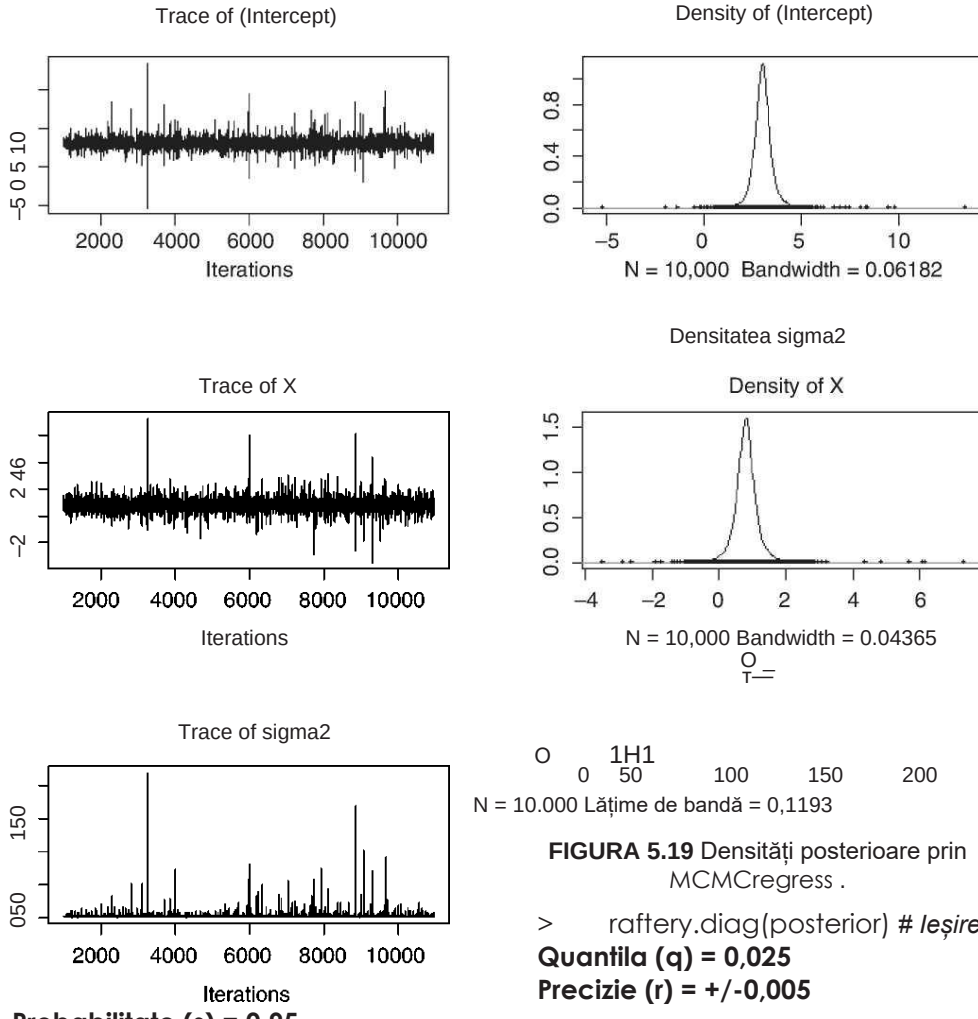


FIGURA 5.19 Densități posterioare prin MCMCregress .

```
>      raftery.diag(posterior) # ieșire:
Quantila (q) = 0,025
Precizie (r) = +/-0,005
Probabilitate (s) = 0,95
```

	Ardere (M)	Total (N)	Mai jos legat (Nmin)	Dependență factorul (I)
(Intercepta)	**3**	**4374**	**3746**	**1.17**
X	**3**	**4374**	**3746**	**1.17**
sigma2	**2**	**3865**	**3746**	**1.03**

> rezumat(posterior) *# Ieșire:*

Iterații = 1001:11000
Interval de subțiere = 1
Număr de lanțuri = 1
Dimensiunea eșantionului per lanț = 10000

1. **Media empirică și abaterea standard pentru fiecare variabilă, plus eroarea standard a mediei:**

	Medie	SD	SE naiv	Seria temporală
(Intercepta)	**3.013**	**0,5459**	**0,005459**	**0,005948**
X	**0,807**	**0,3873**	**0,003873**	**0,004647**
sigma2	**1.510**	**5,0969**	**0,050969**	**0,084921**

2. **Cuantile pentru fiecare variabilă:**

	2,5%	25%	50%	75%	97,5%
(Intercepta)	**2,04394**	**2,7552**	**3,0043**	**3,2483**	**4.083**
X	**0,05681**	**0,6287**	**0,8047**	**0,9769**	**1.562**
sigma2	**0,17091**	**0,3876**	**0,6837**	**1,3395**	**7.155**

>

Modelare în biostatistică

În biostatistică, un *model parametric* este o familie de distribuții care pot fi descrise folosind un număr finit de parametri. Acești parametri sunt de obicei colectați pentru a forma un singur vector parametru *n -dimensional* $q = (q_1, q_2, q_3, \ldots, q_i, \ldots, q_n)$.

Modelele parametrice sunt în contrast cu modelele semiparametrice, semi-neparametrice și neparametrice, toate fiind compuse dintr-un set infinit de „parametri" pentru descriere. Distincțiile dintre aceste patru clase sunt următoarele:

- Într- un *parametric* model, toți parametrii sunt în spații de parametri finiți.
- Un model este *neparametric* dacă toți parametrii sunt în spații de parametri cu dimensiuni infinite.
- Un model *semiparametric conține parametri de* interes cu dimensiuni finite și parametri de neplăcere infiniti.
- Un model *semi-neparametric* are atât parametrii necunoscuți de dimensiuni finite, cât și de dimensiuni infinite de interes.

Dintre instrumentele de modelare disponibile în pachetele CRAN, două importante sunt selectate pentru a fi discutate aici:

1. Pachetul CRAN este potrivit pentru estimarea curbelor doză-răspuns
2. Pachetele CRAN gam (model aditiv generalizat , gam [xi]) de Hastie și Tibshirani (1990) ; și gamair [xi] (Lemn, 2006)

Aceste modele favorizează abordarea *neparametrică* , utilizează tehnici derivate din analiza numerică și teoria aproximării și sunt în general aplicabile tuturor aplicațiilor biostatistice. Ambele sunt ușor adaptabile la calcule folosind R .

PACHETUL CRAN GROFIT PENTRU ESTIMARE CURBELE DOZĂ-RĂSPUNS

Pachetul grofit a fost dezvoltat pentru a se potrivi multor curbe de creștere obținute în diferite condiții pentru a obține o curbă concludentă doză-răspuns. De exemplu, pentru un compus care poate afecta creșterea, grofit potrivește datele la diferite modele parametrice folosind funcția gcFitModel() și oferă, de asemenea, o potrivire spline fără model folosind funcția gcFitSpline () pentru a evita erorile sistematice care ar putea apărea cu aplicația a metodelor parametrice.

De asemenea, în cadrul pachetului grofit , funcțiile R drFitSpline() și drBootSpline() pot fi utilizate pentru a genera un tabel cu estimări pentru jumătatea **concentrației efective maxime** (**EC50**) și statisticile asociate. Termenul *jumătate de concentrație efectivă maximă (* EC50) se referă la concentrația unui medicament, anticorp sau toxic care induce un răspuns la jumătatea distanței dintre valoarea inițială și maximă după un anumit timp de expunere *specificat* . Este folosit în mod obișnuit ca măsură a potenței unui medicament.

Prin urmare, EC50 a unei curbe gradate doză-răspuns reprezintă concentrația unui compus la care se observă 50% din efectul său maxim. EC50 a unei curbe cuantice doză-răspuns reprezintă concentrația unui compus în care 50% din populație prezintă un răspuns, după o durată de expunere specificată. Este, de asemenea, legat de IC50, care este o măsură a inhibării unui compus (inhibare de 50%). Pentru testele de legare a competiției, testele antagoniste funcționale și testele agonist/stimulator, EC50 este cea mai comună măsură rezumativă a curbei doză-răspuns. Răspunsurile la concentrare urmează de obicei o curbă sigmoidală, crescând rapid la o modificare relativ mică a concentrației. Punctul **de inflexiune** la care creșterea eficacității odată cu creșterea concentrației începe să încetinească este EC50. Aceasta poate fi determinată matematic prin derivarea liniei de cea mai bună potrivire. Deși bazarea pe un grafic pentru estimare este mai convenabilă, aceasta produce rezultate mai puțin precise și mai puțin precise.

În cadrul pachetului CRAN grofit , funcția R gcFitModel(x,y) poate fi utilizată pentru sarcini generale de modelare.

PACHELE CRAN gam ȘI gamair

Această subsecțiune prezintă GAM , modelul aditiv generalizat (Hastir & Tibshirani, 1990; Wood, 2006).[22]

Fie Y o variabilă aleatoare de răspuns și X_1 , X_2 , X_3 , …, X_i , …, X_n un set de variabile predictoare. O procedură de regresie poate fi considerată o metodă de estimare a modului în care valoarea lui Y depinde de valorile lui X_1 , …, X_n . Modelul standard de regresie liniară presupune că valoarea așteptată a lui Y are o formă liniară. Valoarea așteptată a lui Y, $E[Y]$ se calculează după cum urmează:

[22] Meulman, JJ *Cursul 4: Modele aditive generalizate.* (2008) . Disponibil la www.math.vu.nl/sto/ onderwijs/statlearn/GAM-OptimalScaling2.pdf

$$= f(X_1, X_2, X_3, \ldots, X_i, \ldots, X_n) \quad (5.2\text{-}37A)$$

$$= b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_i X_i + \ldots + b_n X_n \quad (5.2\text{-}37B)$$

Având în vedere un eșantion de valori pentru Y și X, estimările pentru $b_0, b_1, b_2, b_3, \ldots, b_i, \ldots, b_n$ sunt de obicei obținute prin metoda celor mai mici pătrate.

Modelul aditiv generalizează modelul liniar prin modelarea valorii așteptate a lui Y as

$$E(Y) = f(X_1, X_2, X_3, \ldots, X_i, \ldots, X_n) \quad (5.2\text{-}38A)$$

$$= s_0 + s_1(X_1) + s_2(X_2) + s_3(X_3) + \ldots + s_i(X_i) + \ldots s_n(X_n) \quad (5.2\text{-}38B)$$

unde $s_i(X)$, $i = 1, \ldots, n$ sunt funcții netede. Aceste funcții trebuie estimate într-o abordare *neparametrică* .

GAM extinde modelele liniare tradiționale într-un alt mod, permițând o legătură între $f(X_1, \ldots, X_n)$ și valoarea așteptată a lui Y. Aceasta înseamnă a permite o distribuție alternativă pentru variația aleatoare subiacentă *în afară de distribuția normală* . Deși modelele gaussiene pot fi utilizate în multe aplicații statistice, există tipuri de probleme pentru care nu sunt adecvate. Distribuția normală poate să nu fie adecvată pentru modelarea variabilelor categorice, a răspunsurilor discrete, cum ar fi numărătoarea, sau a răspunsurilor mărginite, cum ar fi proporțiile.

GAM constă dintr-o **componentă aleatorie** , o **componentă aditivă** și o **funcție de legătură** care leagă aceste două componente. Răspunsul Y , componenta aleatoare, se presupune că are o densitate în familia exponențială:

$$f_\gamma(y; q; \varphi) = \exp\{[\{yq - b(q)\}/a(\varphi)] + c(y, \varphi)\} \quad (5.2\text{-}39)$$

unde q se numește *parametrul natural* și φ este *parametrul de scară* . Distribuțiile normale, binomiale și Poisson sunt toate în această familie. Cantitatea

$$h = s_0 + \sum h S_i(X_i) \quad (5,2\text{-}40)$$

unde $s_1(\cdot), \ldots, s_n(\cdot)$ sunt funcții netede, definește componenta aditivă. În cele din urmă, relația dintre media m a variabilei de răspuns și h este definită de o funcție de legătură:

$$g(m) = h \quad (5,2\text{-}41)$$

Funcția de legătură cel mai frecvent utilizată este *legătura canonică* , pentru care

$$h = q \quad (5,2\text{-}42)$$

O combinație de *adaptare* și algoritmi *de scoring local* este utilizată în ajustarea efectivă a modelului.

GAM și GLM pot fi aplicate în situații similare, dar servesc unor scopuri analitice diferite. GLM-urile pun accent pe estimarea și inferența pentru parametrii modelului, în timp ce GAM se concentrează pe explorarea datelor neparametric. Modelul GAM este montat folosind un algoritm de scoring local, care se potrivește iterativ cu modelele aditive ponderate prin adaptare. Algoritmul de backfitting este o metodă Gauss-Seidel pentru adaptarea modelelor aditive, cu netezirea iterativă a reziduurilor parțiale. Algoritmul separă partea parametrică de cea neparametrică a potrivirii, potrivind partea parametrică folosind cele mai mici pătrate liniare ponderate în cadrul algoritmului de adaptare.

Următorul exemplu oferă o introducere elementară în abordarea GAM, folosind R în

calcul.

> **Exemplul 5.25:** O aplicare a GAM la un model liniar simplu de două variabile independente [4]

Fie x_1 și x_2 două variabile predictoare independente. Același tratament poate fi extins la n ($n > 2$) variabile independente. Lasă y să fie variabila rezultată dependentă și e să fie termenul de eroare. Un model liniar poate fi apoi scris ca

$$y = b_1 x_1 + b_2 x_2 + e \tag{5,2-43}$$

Obiectivul modelării este de a minimiza funcția de pierdere în cele mai mici pătrate L definită de

$$L\ (b) = || y b_1 x_1 b_2 x_2 ||^2 \tag{5,2-44}$$

Un simplu GAM caută modelul dat de

$$L\ [f(x)] = || y f_1(x_1) - f_2(x_2) ||^2 \tag{5,2-45}$$

unde $f_1(x_1)$ și $f_2(x_2)$ sunt funcții – în special transformări neliniare ale x_1 și x_2.

Calculul acestor funcții, $f_1(x_1)$ și $f_2(x_2)$, poate fi acum efectuat folosind următorul segment de cod R :

```
> n <- 100 # Luând în considerare 100 de puncte de date
> x1 <- scară(runif(n, 0, 1))
> x2 <- scară(runif(n, 0, 1))
> # Funcția runif(n, min=0, max=1) oferă informații despre
> # distribuția uniformă pe intervalul de la min la max.
> # runif() generează abateri aleatorii.
> # Distribuția uniformă are densitate
> # f(x) = 1/(max - min)
> # pentru min ≤ x ≤ max.
> # scale() este o funcție generică a cărei metodă implicită se centrează și/sau se scalează
> # coloanele unei matrice numerice.

> f <- x1**2 + x2**2
> e <- rnorm(n, 0, 0,1)
> # Funcția rnorm(n, mean, sd) generează aleatoriu în mod normal
> # numere distribuite cu media egală cu media și abaterea standard
> # egal cu sd.
>
> y <- f + e
> a <- lm(y ~ cbind(x1,x2))
> yhat <- a$fited.values
> mr2 <- cor(yhat,y)**2
> plot(yhat, y, main=c('multiplu R2=',(round(mr2,digit=5))))
> # Ieșire: Figura 5.20 .
```

Multiplu R2 = 0,00716

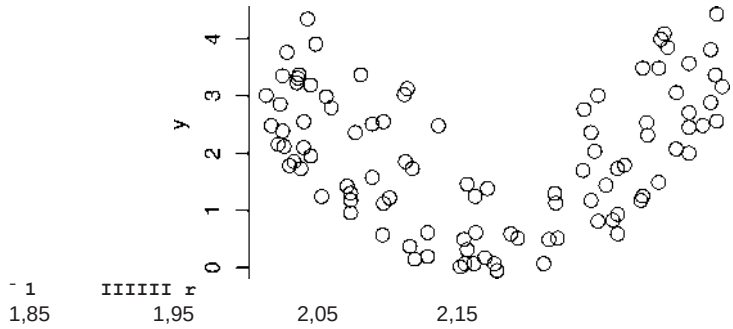


FIGURA 5.20 Model liniar.

```
>  # Acum, pentru GAM, următorul calcul va genera funcțiile
>  # f1(x1) și f2(x2):
>  # GAM pas cu pas:
>  xx1 <- x1
>  xx2 <- x2
>
>  plot((x1+x2),y, main='y ca sumă a x1 și x2')
>  # Ieșire Figura 5.21 GAM-1
>  plot(x1,(y-x2), main='(y - x2) versus x1')
>  # Ieșire: Figura 5.22 GAM-2
>  dev.copy2eps(file="plots1.eps")

> plot(xx1,(y-xx2),main='(y - f2(x2)) versus f1(x1)')
>  # Ieșire: Figura 5.23 GAM-3
```

```
>   plot(xx2,(y-xx1),main='(y - f1(x1)) versus f2(x2)')
>   # Ieșire: Figura 5.24 GAM-4
> plot((xx1+xx2),y-(xx1+xx2),asp=1,principal='Reziduuri')
>   # Ieșire: Figura 5.25 GAM-5
>   plot(x1,xx1,main='Transformarea lui x1')
>   linii(x1[comanda(x1)],xx1[comanda(x1)],col='roșu')
>   # Ieșire: Figura 5.26 GAM-6
>   plot(x2,xx2,main='Transformarea lui x2')
> linii(x2[comanda(x2)],xx2[comanda(x2)],col='albastru')
>   # Ieșire: Figura 5.27 GAM-7
```

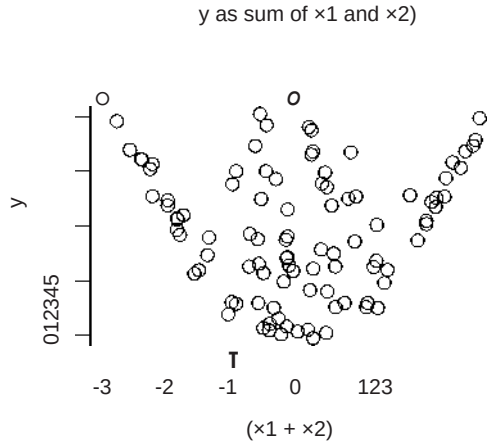


FIGURE 5.21 GAM-1.

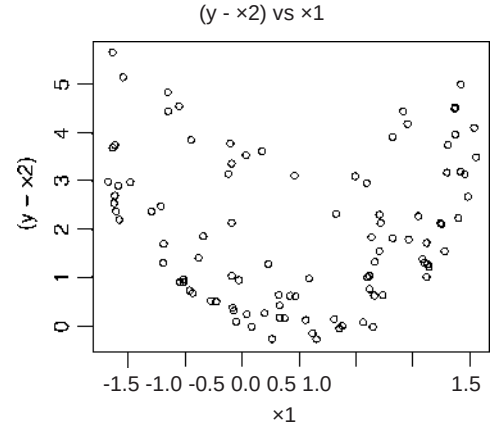


FIGURE 5.22 GAM-2.

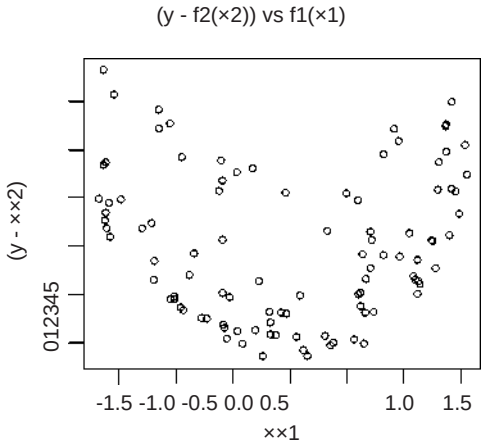


FIGURE 5.23 GAM-3.

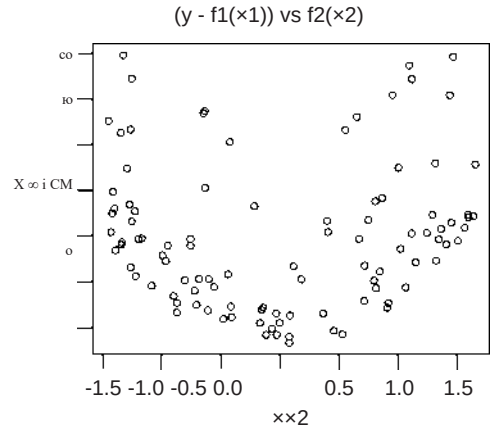


FIGURE 5.24 GAM-4.

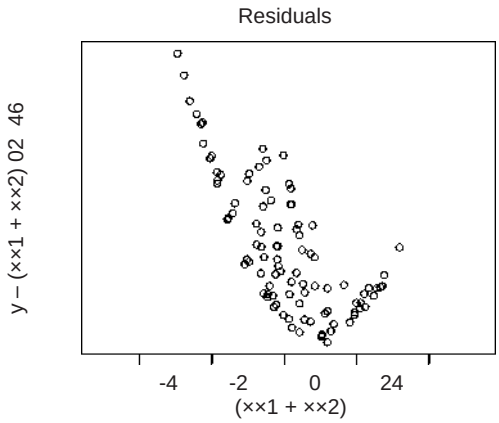


FIGURE 5.25 GAM-5.

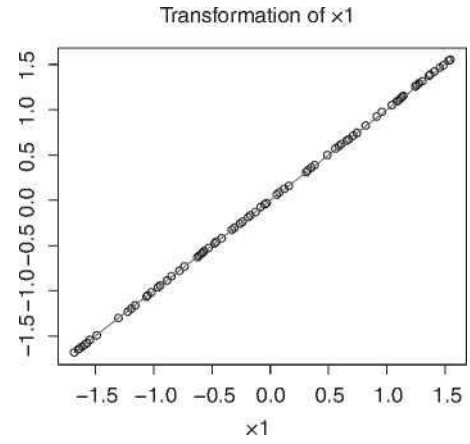


FIGURE 5.26 GAM-6.

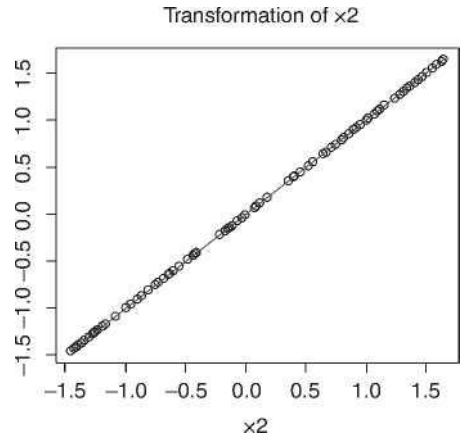


FIGURE 5.27 GAM-7.

Întrebări de revizuire pentru Secțiunea 5.2

1. (a) Ce este biostatistica bayesiană?
   (b) Exprimați teorema lui Bayes în termeni probabilistici.
2. (a) Ce este inferența bayesiană?
   (b) Exprimați inferența bayesiană în termeni probabilistici.
3. (a) Ce este analiza supraviețuirii?
   (b) Explicați pe scurt următorii termeni: *timp de supraviețuire* , *date de supraviețuire* , *timp de supraviețuire cenzurat* și *date cenzurate* .
4. (a) Care este procedura K–M în analiza supraviețuirii?
   (b) Folosind teoria probabilității, deduceți o expresie pentru probabilitatea ca un subiect de caz să supraviețuiască până la timpul t .
5. Într-un grafic tipic K–M al subiecților cazului supraviețuitori, ceea ce este reprezentat de
   (a) un segment vertical?
   (b) un segment orizontal?
   (c) +?

6. (a) În analiza supraviețuirii, ce este modelul de regresie Cox (cunoscut și sub denumirea de *pericole proporționale*)?
   (b) În modelul de regresie Cox, ce se înțelege prin *rata de eșec instantanee* , *funcția hazardului de referință* și *regresia hazardelor proporționale* ?
7. În procedura K–M, explicați utilizările următoarelor funcții R :
   (a) Surv()
   (b) survfit()
   (c) survdiff()
   (d) surv_test()
8. În modelul de regresie Cox, explicați utilizările următoarelor funcții R :
   (a) survfit()
   (b) coxph()
   (c) cox.zph()
9. (a) Ce se înțelege prin formula Bayes inversă (IBF)?
   (b) Care sunt avantajele utilizării IBF în modelarea biostatistică?
10. (a) În pachetul CRAN grofit , care a fost dezvoltat pentru potrivirea curbelor de creștere, explicați utilizările următoarelor funcții R : gcFitModel() , gcFitSpline() și drBootSpline().
    (b) În pachetul CRAN gam , care a fost dezvoltat pentru modelarea aditivă generalizată , explicați utilizările funcției R gam () și utilizarea unui gam.object.

Exerciții pentru Secțiunea 5.2

1. Un studiu realizat de Gehan, raportat de Daniel (2005) , a încercat să găsească doza optimă a medicamentului analgezic lignocaină (LC), care a fost introdus prin injectare cu propofol (PF) subiecților cazați. Un total de 310 subiecți caz au fost implicați, iar acești subiecți caz au fost alocați în patru categorii în funcție de doza LC:
Grupa A = subiecți care nu au primit LC
Grupele B, C și D = subiecți care au primit 0,1, 0,2 și, respectiv, 0,4 mg/kg LC amestecat cu PF
Gradul de durere experimentat de subiecții cazului a fost clasificat de la 0 (fără durere) la 3 (cel mai sever). Următorul tabel înregistrează subiecții de caz clasificați încrucișați după nivelul de dozare și scorul durerii:

		Grup				Total
	Scorul durerii	O	B	C	D	
	0	49	73	58	62	242
	1	16	7	7	8	38
	2	8	5	6	6	25
	3	4	1	0	0	5
Total	77	77	86	71	76	310

(a) Calculați următoarele probabilități și explicați semnificația lor:
(i) $P(2 \cap A)$

(ii) P (B ∪ 3)
(iii) P (0 | C)

(b) Este adevărată fiecare dintre următoarele ecuații? De ce sau de ce nu?
(i) P (A ∩ 3) = P (3 ∩ A)
(ii) P (1 ∪ D) = P (D и 1)
(iii) P (в ∩ D) = 0

2. Într-un studiu clinic, subiecții cazului sunt împărțiți în două grupuri (0 = control, ■ = Tratament).[23] Probabilitatea ca un rezultat advers să apară în grupul de control este P 0, iar în grupul de tratament este P 1. Acești subiecți de caz sunt plasați alternativ în cele două grupuri, iar rezultatele lor sunt independente. Folosind teoria probabilității, teorema lui Bayes sau altă formulă, arată că probabilitatea ca primul eveniment advers să apară în grupul de control este

$$P0/(\ P0 + P\ 1 - P\ O \times P\ 1)$$

Dovada:
Defini:
- Spațiul de probă să fie format din toate secvențele infinite posibile de rezultate ale pacientului
- Evenimentul E1 – primul subiect (alocat brațului de control) suferă un rezultat advers
- Evenimentul E2 – primul subiect (alocat brațului de control) nu suferă un - rezultat advers, dar al doilea pacient (alocat brațului de tratament) suferă un rezultat advers
- Evenimentul E0 – niciunul dintre primii doi subiecți nu suferă un rezultat advers
- Evenimentul F – primul eveniment advers are loc la brațul de control

Pentru a răspunde la această întrebare, trebuie să găsiți P (F).

Evenimentele E1, E2 și E0 împart, deci, prin teorema probabilității totale, P (F) = P (F|E1) P (E1) + P (F|E2) P (E2) + P (F| E0) P (E0). De asemenea, P (E1) = P 0, P (E2) = (1 - P 0) P 1, P (E0) = (1 - P 0) (1 - P 1) și P (F|E1) = 1, P (F|E2) = 0.

În cele din urmă, deoarece după două rezultate nedefavorabile procesul de alocare repornește efectiv, unul are P (F|E0) = P (F). Prin urmare:

$$P\ (F) = (1 \times P\ 0) + (0 \times (1 - P\ 0)\ P\ 1) + (P(F) \times (1 - P\ 0)\ (1 - P\ 1))$$

$$= P\ 0 + (1 - P\ 0)\ (1 - P\ 1)\ P(F),$$

care poate fi rearanjat pentru a da P (F) = P 0/(P 0 + P 1 - P 0 P 1), după cum este necesar.

3. *Procedura K–M folosind* R .

Folosind următorul segment de cod R , procedura K–M poate fi aplicată setului de date melanom din pachetul ISwR :

```
> install.packages(„supraviețuire")
> bibliotecă (supraviețuire)
> ls(„pachet:supraviețuire")
> install.packages ("ISwR")
```

[23] Exemple lucrate *1: Probabilitatea totală și teorema lui Bayes.* Disponibil la www2.imperial.ac.uk/~ayoung/ m2s1/WorkedExamples1.pdf

```
> bibliotecă (ISwR)
> ls("pachet:ISwR")
> date (melanom)
> atașează (melanom)
> melanom
> nume (melanom)
> survfit(Surv(zile, stare==1)~1,date=melanom)
> plot(survfit(Surv(zile,starea==1)~1,data=melanom))
> # ieșire: Figura 5.28 .
> rezumat(survfit(Surv(zile, stare==1)~1,date=melanom))
```

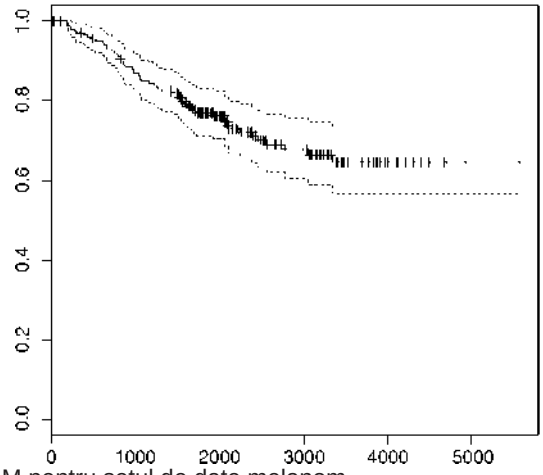


FIGURA 5.28 Graficul K–M pentru setul de date melanom.

> a) Descrieți, cu propriile cuvinte, semnificația și scopul fiecăreia dintre comenzile R din segmentul de cod precedent.
> b) Sugerați comenzi R alternative care ar putea obține aceleași rezultate.
> c) Introduceți mediul R și rulați segmentul de cod pentru a obține Figura 5.28 .
> d) Sugerați comenzi R pentru a adăuga etichete la axele și un titlu potrivit pentru Figura 5.28 .

4. *Modelul de regresie Cox (riscuri proporționale) folosind* R *.*

Folosind următorul segment de cod R , procedura modelului de regresie Cox poate fi aplicată setului de date privind cancerul din pachetul de supraviețuire :

```
> install.packages(„supraviețuire")
> bibliotecă (supraviețuire)
> ls(„pachet:supraviețuire")
> date (cancer)
> atașa (cancer)
> cancer
> lfit6 <- survreg(Surv(time, status)~pspline(varsta, df=2), cancer)
> plot(cancer$age, predict(lfit6), xlab="Vârsta", ylab="Spline
+ predicție")
> titlu ("Date despre cancer")
```

```
>   # Ieșire: Figura 5.29 .
> fit0 <- coxph(Surv(time, status) ~ ph.ecog + age, cancer)
> fit1 <- coxph(Surv(time, status) ~ ph.ecog + pspline(age,3),
+                 cancer)
> fit3 <- coxph(Surv(time, status) ~ ph.ecog + pspline(varsta,8),
+                 cancer)
> fit0
>   potrivi1
>   potrivi3
```

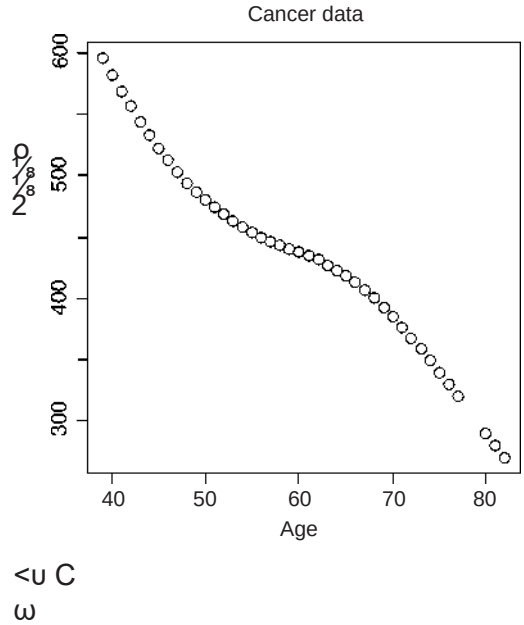


FIGURA 5.29 Modelul de regresie Cox pentru setul de date despre cancer .

(a) Descrieți, cu propriile cuvinte, semnificația și scopul fiecăreia dintre comenzile R din segmentul de cod precedent.
(b) R alternative care ar putea obține aceleași rezultate.
(c) Introduceți mediul R și rulați segmentul de cod pentru a obține Figura 5.29 .
(d) Urmând comanda > ls("pachet:survival") : Setul de date cancer este inclus în pachetul de supraviețuire ?
(e) Urmând comanda > cancer : Câte cazuri sunt incluse în acest set de date?

5. *Mai multe despre procedura K–M folosind* R .

Folosind următorul segment de cod R , procedura K–M poate fi aplicată setului

de date cardiace din pachetul CRAN de supraviețuire . (Înainte de a genera diagrame, se efectuează o preprocesare pentru a obține acest set de date în forma adecvată pentru eveniment. Funcția de istorie. Trebuie să creați o linie per subiect și să sortați după timp sub observație, cei care se confruntă cu un eveniment înaintea celor legați de timpul de cenzură. .)

```
>  install.packages(„supraviețuire")
>  bibliotecă (supraviețuire)
>  ls(„pachet:supraviețuire")
>  date (inima)
>  inima

>  # Crearea versiunii event.history a setului de date inimii ( numită heart.one):
>  heart.one <- matrix(nrow=length(unique(heart$id)), ncol=8)
>  pentru(i în 1:lungime(unic(id-inima$)))

+ {
+ if(lungime(heart$id[heart$id==i]) == 1)
+ heart.one[i,] <- as.numeric(unlist(heart[heart$id==i, ]))
+ else if(lungime(id$id inimii[id$id==i]) == 2)
+ heart.one[i,] <- as.numeric(unlist(heart[heart$id==i,][2,]))
+ }

>  heart.one[,3][heart.one[,3] == 0] <- 2

>  # Conversia evenimentelor cenzurate la 2, de la 0
>  daca(este.factor(inima$transplant))
+ heart.one[,7] <- heart.one[,7] - 1

>  ## Revenirea la codificarea corectă a transplantului
>  inima.unul <
+     ca.data.frame(heart.one[order(unlist(heart.one[,2]),
+     unlist(heart.one[,3])),])
>  nume(inima.unu) <- nume(inima)

>  # Înapoi la indicatorul obișnuit de cenzură:
>  heart.one[,3][heart.one[,3] == 2] <- 0

>  # Notă: transplantul spune 0 (pentru niciun transplant) sau 1 (pentru un transplant)
>  # și evenimentul = 1 este moartea, în timp ce evenimentul = 0 este cenzurat .
>  # Trasează o singură curbă K–M din datele inimii , mai întâi creând
>  # un obiect de supraviețuire
>  heart.surv <- survfit(Surv(stop, event) ~ 1, data=heart.one,
+                       conf.int = FALSE)
>  # Curba K–M tradițională
>  # postscript('ehgfig3.ps', horiz=TRUE)
>  # omi <- par(omi=c(0,1.25,0.5,1.25))
>  plot(heart.surv, ylab='probabilitate estimată de supraviețuire',
+                  xlab='timpul de observație (în zile)')
>  title('Curba Kaplan-Meier pentru datele Stanford', cex=0,8)

>  Ieșire: Figura 5.30 .
```

(a) Descrieți, cu propriile cuvinte, semnificația și scopul fiecăreia dintre comenzile R din segmentul de cod precedent.
(b) R alternative care ar putea obține aceleași rezultate.
(c) Introduceți mediul R și rulați segmentul de cod pentru a obține Figura 5.30 .

6. *Mai multe despre modelul de regresie Cox folosind* R *.*

Folosind următorul segment de cod R , procedura modelului de regresie Cox poate fi aplicată setului de date veteran despre cancer, utilizând programul plot.cox.zph în

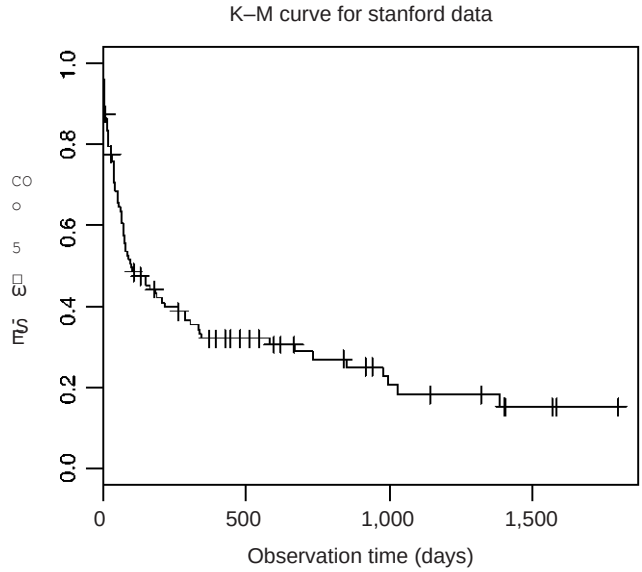


FIGURE 5.30 K–M survival plot for the cancer dataset.

supraviețuirea pachetului CRAN . Programul afișează un grafic al reziduurilor de feld Schoen la scară , împreună cu o curbă netedă, dezvoltată pentru un test grafic al pericolelor proporționale :

```
>  install.packages(„supraviețuire")
>  bibliotecă (supraviețuire)
>  ls(„pachet:supraviețuire")
>  date (veteran)
>  atașați (veteran)
>  veteran
>  vfit <- coxph(Surv(time,status) ~ trt + factor(celltype) +
+                  karno + vârstă, date=veteran, x=ADEVĂRAT)
>  temp <- cox.zph(vfit)
> plot(temp, var=5) # Uită-te la scorul Karnofsky; acesta este un mod vechi de a
  face intriga.
> plot(temp[5]) # Mod nou cu subscripting:
>  abline(0, 0, lty=3)
>  # Adăugați și potrivirea liniară:
>  abline(lm(temp$y[,5] ~ temp$x)$coeficienți, lty=4, col=3)
>  titlu(principal="Studiu pulmonar VA")
```

> *Ieșire:* Figura 5.31 .

(a) Descrieți, cu propriile cuvinte, semnificația și scopul fiecăreia dintre comenzile R din segmentul de cod precedent .

(b) R alternative care ar putea obține aceleași rezultate.

(c) Introduceți mediul R și rulați segmentul de cod pentru a obține Figura 5.31 .

(d) Urmând comanda > ls("pachet:survival") : veteranul setului de date este inclus în pachetul de supraviețuire ?

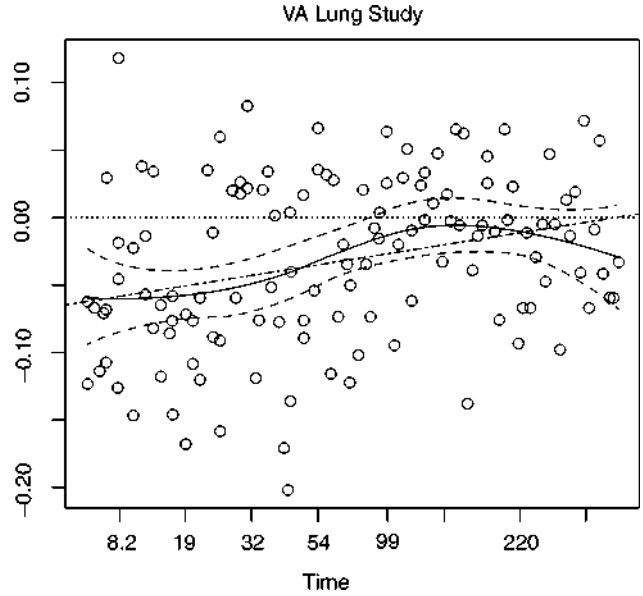


FIGURA 5.31 Modelul de regresie Cox pentru setul de date de cancer modificat .

(e) Urmând comanda > veteran : Câte cazuri sunt incluse în acest set de date?

(f) Care este linia verde din figura 5.31 ?

7. *Mai multe despre analiza supraviețuirii: Modelul de regresie Cox, bazat pe reziduuri cumulate , folosind* R .

Folosind următorul segment de cod R , aplicați procedura modelului de regresie Cox , bazată pe reziduurile cumulate, la setul de date pbc [ciroza biliară primară (PBC) sau ciroza ficatului] Clinica Mayo. Aceste date au fost obținute din studiile clinice la subiecții cu CBP a ficatului, efectuate între 1974 și 1984. CBP este o boală autoimună a ficatului marcată de distrugerea lentă, progresivă a căilor biliare mici din ficat. Când aceste canale sunt deteriorate, bila se acumulează în ficat și, în timp, dăunează țesutului, ceea ce poate duce la cicatrizare, fibroză și ciroză. *Ciroza* este o consecință a bolii hepatice cronice caracterizată prin înlocuirea țesutului hepatic cu fibroză, țesut cicatricial și noduli regenerativi (bulgări care apar ca urmare a unui proces în care țesutul deteriorat este regenerat), ducând la pierderea funcției hepatice. Ciroza este cauzată cel mai frecvent de alcoolism, hepatită B și C și boala ficatului gras, dar are și multe alte cauze posibile. Unele cazuri sunt idiopatice, (adică, de cauză necunoscută). Studii recente au arătat că poate afecta până la 1 din 3.000 de persoane, iar raportul de gen este de cel puțin 9:1 (femeie/bărbați).

Un total de 424 de pacienți cu PBC, toți trimiși la Clinica Mayo în acel interval

de 10 ani, au îndeplinit criteriile de eligibilitate pentru studiul randomizat controlat cu placebo al medicamentului D-penicilamină. Primele 312 cazuri din setul de date au participat la studiul randomizat, iar datele pentru aceste cazuri sunt în mare parte complete. Cele 112 cazuri suplimentare nu au participat la studiul clinic, dar au consimțit să li se înregistreze măsurătorile de bază și să fie urmărite pentru supraviețuire. Șase dintre aceste cazuri au fost pierdute

urmărire la scurt timp după diagnostic, astfel că datele de aici sunt pentru încă 106 cazuri, precum și pentru cei 312 participanți randomizati, însumând 106 + 312 = 418 cazuri:

```
> install.packages(„gof")
> bibliotecă (gof)
> ls("pachet:gof")
> cumres
> exemplu (cumres)
> install.packages(„supraviețuire")
> bibliotecă (supraviețuire)
> ls(„pachet:supraviețuire")
> date (pbc)
> atașează (pbc)
> pbc
> ## Exemplu PBC
> fit.cox <- coxph(Surv(time,status==2) ~ varsta + edem + bili +
>                                                protime, data=pbc)
> system.time(pbc.gof <- cumres(fit.cox,R=2000))
> par(mfrow=c(2,2))
> plot(pbc.gof, ci=TRUE, legend=NULL)
> # ieșire: Figura 5.32 .
```

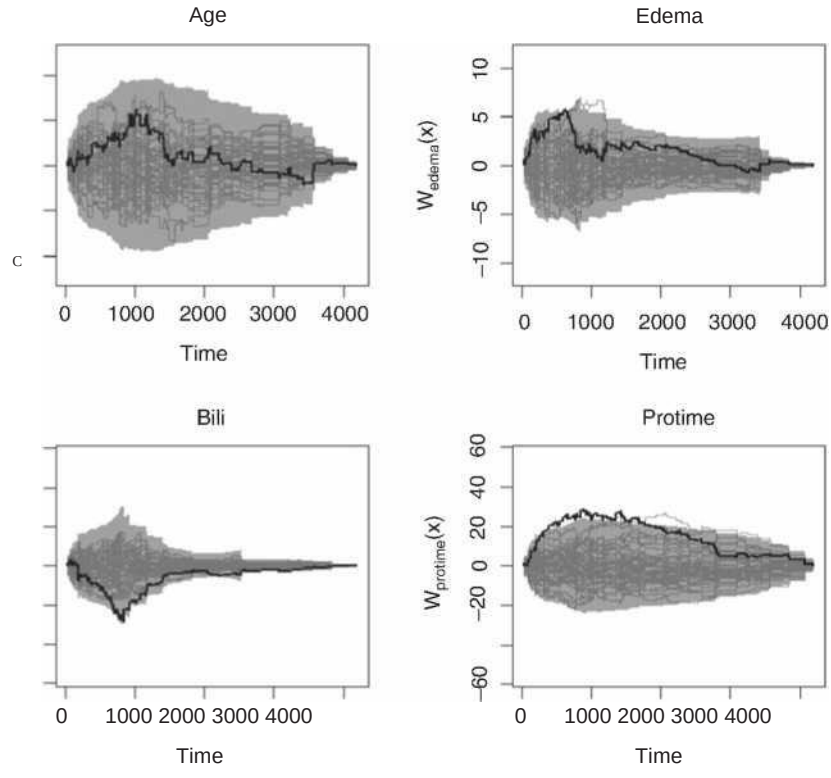


FIGURE 5.32 Survival modeling of the pbc dataset: the Cox regression model, based on cumulative residues.

(a) Descrieți, cu propriile cuvinte, semnificația și scopul fiecăreia dintre comenzile R din segmentul de cod precedent.
(b) Sugerați comenzi R alternative care ar putea obține aceleași rezultate.
(c) Introduceți mediul R și rulați segmentul de cod pentru a obține Figura 5.32 .
(d) Urmând comanda > ls("pachet:survival") : Setul de date pbc este inclus în pachetul survival ?
(e) Urmând comanda > pbc : Câte cazuri sunt incluse în acest set de date?

8. *Un exemplu de aplicare a formulei Bayes inversă* (Ng & Tong, 2010)

O critică la adresa tehnicii de simulare Monte Carlo pentru generarea de eșantioane aleatoare din distribuții univariate și multivariate se referă la momentul respingerii unei selecții. O metodă recunoscută este propunerea Rubin a unei proceduri de eșantionare neiterativă: metoda de **eșantionare/resampling de importanță (SIR)** . Pentru simularea dintr-o densitate definită în intervalul unitar, fie *r* un număr întreg pozitiv cunoscut și $X \sim f(x)$, unde:

$$f(x) = p \sin^r(px) / B[½, (r+1)/q], \quad 0 < x < 1 \qquad (5,2\text{-}46)$$

iar *B* este funcția beta.

Bazat pe Ecuația (5.2-46) , următorul segment de cod R calculează $f(x)$ când $r = 6$, pentru o densitate beta oblică $B(x|2, 4)$, ca densitate de eșantionare importantă $g(x)$. Figura 5.33 (A) arată ponderea importanței $w(x) = f(x)/B(x|2, 4)$. Algoritmul - stabilește $J = 200.000$ și $m = 20.000$ pentru a obține Figura 5.33 (B), care arată că histograma recuperează în întregime funcția de densitate țintă $f(x)$.

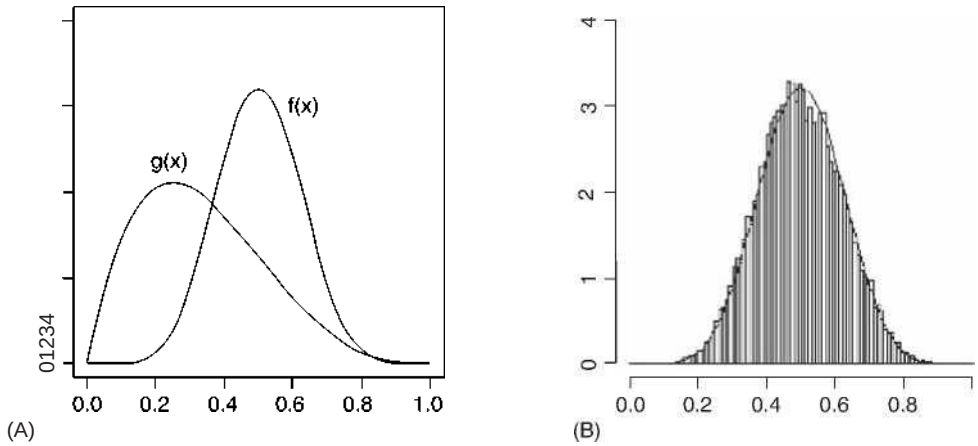


FIGURE 5.33 Output of the SIR method for noniterative sampling.

Iată segmentul de cod R necesar :

```
> J = 200000
> m = 20000
> # Numele funcției: IBF2.2.SIR(J=200.000, m=20.000)
> # ********** Input **********************************
> # J = 200,000 is the sampling size in the SIR method
> # m = 20,000 is the resampling size in the SIR method
> # ************************************************
>
>  # Scop: Complot Figurile 5.33(A) și 5.33(B)
```

```
> r <- 6
> a <- 2
> b <- 4
> x <- secv(0, 1,0.01) > x
```

```
[1]  0,00  0,01  0,02  0,03  0,04  0,05  0,06  0,07  0,08  0,09  0,10  0,11  0,12  0,13
0,14
[16] 0,15  0,16  0,17  0,18  0,19  0,20  0,21  0,22  0,23  0,24  0,25  0,26  0,27  0,28
0,29
[31] 0,30  0,31  0,32  0,33  0,34  0,35  0,36  0,37  0,38  0,39  0,40  0,41  0,42  0,43
0,44
[46] 0,45  0,46  0,47  0,48  0,49  0,50  0,51  0,52  0,53  0,54  0,55  0,56  0,57  0,58
0,59
[61] 0,60  0,61  0,62  0,63  0,64  0,65  0,66  0,67  0,68  0,69  0,70  0,71  0,72  0,73
0,74
[76] 0,75  0,76  0,77  0,78  0,79  0,80  0,81  0,82  0,83  0,84  0,85  0,86  0,87  0,88
0,89
[91] 0,90  0,91  0,92  0,93  0,94  0,95  0,96  0,97  0,98  0,99  1,00
```

```
>
> cc <- (gamma(0,5) * gamma((r + 1)/2))/gamma(0,5 * r + 1)
> fx <- (pi * (sin(pi * x))^r)/cc
> gx <- dbeta(x, a, b)
> #-------------------------------------------------------
> xJ <- rbeta(J, a, b)
> w <- (sin(pi * xJ))^r/dbeta(xJ, a, b)
> p <- w/sum(w)
> xsample <- eșantion(xJ, m, prob = p, înlocuiți = F)
> # ---------- (a)----------
> par(pty = "s")
> par(mfrow = c(1,2))
> plot(x, fx, tip = "l", lty = 1, ylim = c(0, 4), xlab = "",
+ ylab = "", main = "(a)")
> linii (x, gx, tip = "l", lty = 4)
> text(0,65, 3, „f(x)", cex = 1,8)
> text(0,22, 2,5, „g(x)", cex = 1,8)
> # ieșire: Figura 5.33 (A).
> # ---------- (b)----------
> hist(xsample, probabilitate = T, ylim = c(0, 4), xlab = " ",
+ pauze = seq(0, 1,0.01), ylab = "", main = "(b)")
> lines(x, fx, type = "l", lty = 1) # ieșire: Figura 5.33 (B).
```

(a) Densitatea țintă $f(x)$, așa cum este definită de ecuația (5.2-46) , cu $r = 6$ și $g(x) = B(x|2, 4)$.
(b) Histograma lui $f(x)$ obținută prin metoda SIR, cu $J = 200.000$ și $m = 20.000$.
(c) Descrieți, cu propriile cuvinte, semnificația și scopul fiecăreia dintre comenzile R din segmentul de cod precedent.
(d) Sugerați comenzi R alternative care ar putea obține aceleași rezultate.
(e) Introduceți mediul R și rulați segmentul de cod pentru a obține Figura 5.33 .

(f) Repetați calculul pentru diferite valori ale lui J (mai mare și mai mică de 200.000) și m (mai mare și mai mică de 20.000).

(g) Cum diferă aceste rezultate noi de rezultatele primului calcul? Ex clar.

9. *Aplicarea funcției* cu() în baza *pachetului CRAN .*

Funcția cu() evaluează o expresie R într-un mediu construit din date, eventual modificând datele originale. Începând cu pachetul gamair , [1] exemplul următor folosește un segment scurt de cod R pentru a ilustra aplicabilitatea simplă a modelării jocurilor (din pachetul gamair) . gam este folosit pentru a modela temperatura medie a aerului (în grade Fahrenheit) în Cairo, Egipt, de la 1 ianuarie 1995 până la 21 mai 2005 (Wood, 2006) .

Utilizare:
date (cairo)

Format:
Un cadru de date cu 6 coloane și 3.780 de rânduri. Coloanele sunt:

- lună — Luna anului de la 1 la 12
- ziua.lunii — Ziua lunii, de la 1 la 31
- an — Anul, începând cu 1995
- temp — Temperatura medie (grade Fahrenheit)
- ziua.a.anului — Ziua anului de la 1 la 366
- timp — Numărul de zile de la 1 ianuarie 1995

Sursa: Wood (2006) , pp. 321–324.

Calculul R se bazează pe date(cairo) cu instrucțiunea:
(cairo,plot(time,temp,type="l")

Se poate utiliza următorul segment de cod R :

```
> install.packages("gamair") # Se instalează pachetul gamair
> bibliotecă (gamair)
> ls("pachet:gamair")
caracter (0)
> date (cairo)
> # Autorul vizita acest oraș egiptean antic în momentul scrierii > # prima schiță a acestei porțiuni a cărții!
> # Sursa: http://www.engr.udayton.edu/weather/citylistWorld.htm
> atașa (cairo)
> cairo # Inspectarea cadrului de date cairo
> # ieșire: un cadru de date cu 6 coloane și 3780 de rânduri.
```

	lună	ziua.lunii	an	temp	ziua.a.anului	timp
1	1	1	1995	59.2	1	1
2	1	2	1995	57,5	2	2
3	1	3	1995	57.4	3	3
4	1	4	1995	59.3	4	4

5	**1**		**5**	**1995**	**58,8**	**5**	**5**
3790	**5**	**17**	**2005**	**78.1**	**137**	**3790**
3791	**5**	**18**	**2005**	**79,9**	**138**	**3791**
3792	**5**	**19**	**2005**	**82,7**	**139**	**3792**
3793	**5**	**20**	**2005**	**83,5**	**140**	**3793**
3794	**5**	**21**	**2005**	**76,9**	**141**	**3794**

```
>
> # Funcția cu():
> ls("pachet:bază")
```

[1] „-” "-.Data" "-.POSIXt" "!"
[5] „!.hexmode” "!.octmode" "!=" "$"

[1169] „în timp ce” „cu” „cu.default” „cuCallingHandlers”

[1197] „xtfrm.Surv” „xzfile” „zapsmall”

```
> # Afișează 1199 de fișiere
> cu
```

funcție (date, expr, ...)
UtilizațiMetoda(„cu”)
<bytecode: 06DB2644 >
<mediu: namespace:base >

```
> cu(cairo, plot(time,temp,type="l"))
> # Ieșire: Figura 5.34 .
```

(a) Descrieți, cu propriile cuvinte, semnificația și scopul fiecăreia dintre comenzile R din segmentul de cod precedent.

(b) R alternative care ar putea obține aceleași rezultate.

(c) Introduceți mediul R și rulați segmentul de cod pentru a obține Figura 5.34 .

(d) Urmând comanda > ls("pachet:bază") :

(i) Funcția cu () este inclusă în baza pachetului ?

(ii) Câte fișiere sunt incluse în acest pachet?

(e) *Mai multe despre* gam *(modele aditive generalizate): Funcția* R vis.gam() *și vizualizarea obiectelor* gam *.*

Pachetul CRAN mgcv (gcv = validare încrucișată generalizată) conține funcții pentru modelarea aditivă generalizată și modelarea mixtă aditivă generalizată. Termenul *GAM* este considerat ca include orice GLM estimat prin maximizarea probabilității (posibil cvasi) penalizată pătratic. Este oferită o abordare bayesiană a calculului intervalului de încredere/credibil. Funcționale liniare (adică, funcțiile funcțiilor) ale netezirii, penalizarea termenilor modelului parametric și legarea parametrilor de netezire sunt toate acceptate.

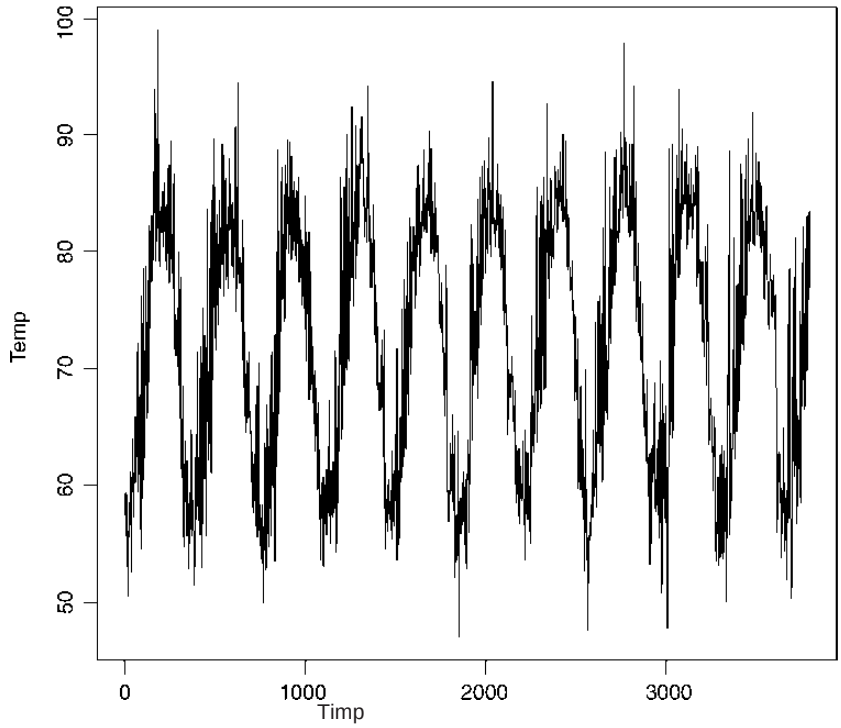


FIGURA 5.34 Modelul de aditiv generalizat gamair() pentru datele de temperatură zilnică pentru Cairo, Egipt.

Luați în considerare următorul segment de cod R :

```
>  install.packages("mgcv")
>  bibliotecă (mgcv)
>  ajutor(„pachet-mgcv")
>  ls("pachet:mgcv")
>  set.seed(0)
>  n <- 200; sig2 <- 4
>  x0 <- runif(n, 0, 1); x1 <- runif(n, 0, 1)
>  x2 <- runif(n, 0, 1)
>  y <- x0^2 + x1*x2 + runif(n,-0,3,0,3)
>  g <- gam(y~s(x0,x1,x2))
>  vechi.par <- par(mfrow=c(2,2))
>  # afișați suprafața de predicție în x0, x1 .....
>  vis.gam(g,ticktype="detailed",color="heat",theta=-35)
>  vis.gam(g,se=2,theta=-35)
>  # cu suprafețe de eroare standard de două ori
>  vis.gam(g, view=c("x1","x2"),cond=list(x0=0.75))
>  # vedere diferită
>  vis.gam(g, view=c("x1","x2"),cond=list(x0=.75), theta=210,
+       phi=40, de asemenea.far=.07)
```

> # *Ieșire:* Figura 5.35 .

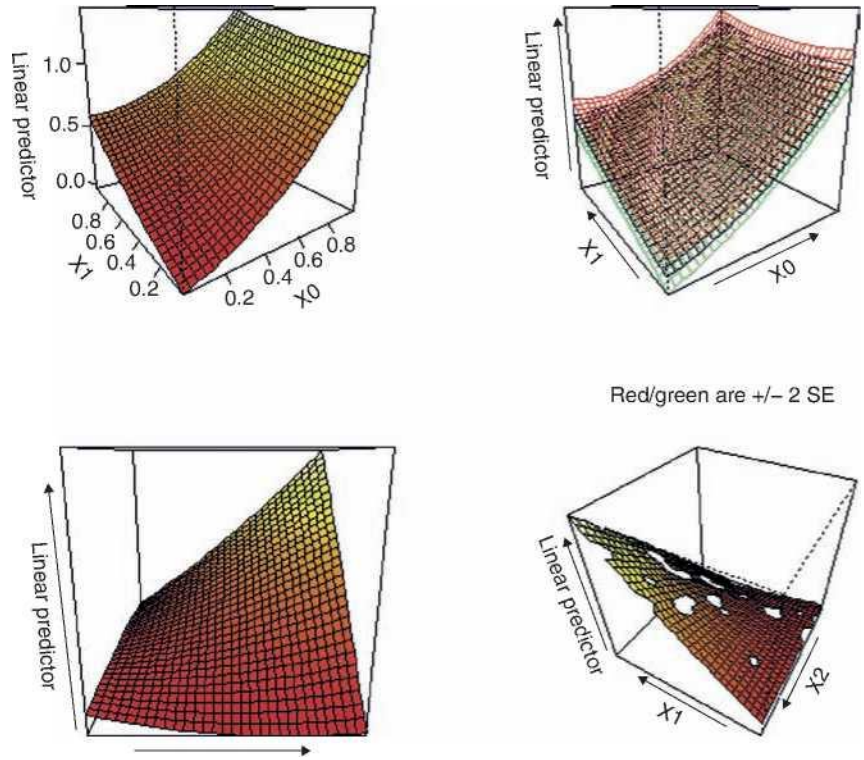


FIGURA 5.35 Diagrame de contur ale suprafețelor de predicție și erori standard folosind vis.gam() .

```
>
>   # . zone în care nu există date nu sunt reprezentate grafic
>
>   # exemple de contur ....
>   vis.gam(g, view=c("x1","x2"),plot.type="contour",color="heat")
>   vis.gam(g, view=c("x1","x2"),plot.type="contour",
+             culoare="terren")
>   vis.gam(g, view=c("x1","x2"),plot.type="contour",color="topo")
>   vis.gam(g, view=c("x1","x2"),plot.type="contour",color="cm")
>   # Ieșire: Figura 5.36 .
>
>   par(vechi.par)
>   # Exemple cu variabile factor și „de” :
>   fac <- rep(1:4,20)
>   x <- runif(80)
>   y <- fac+2*x^2+rnorm(80)*0.1
>   fac <- factor(fac)
```

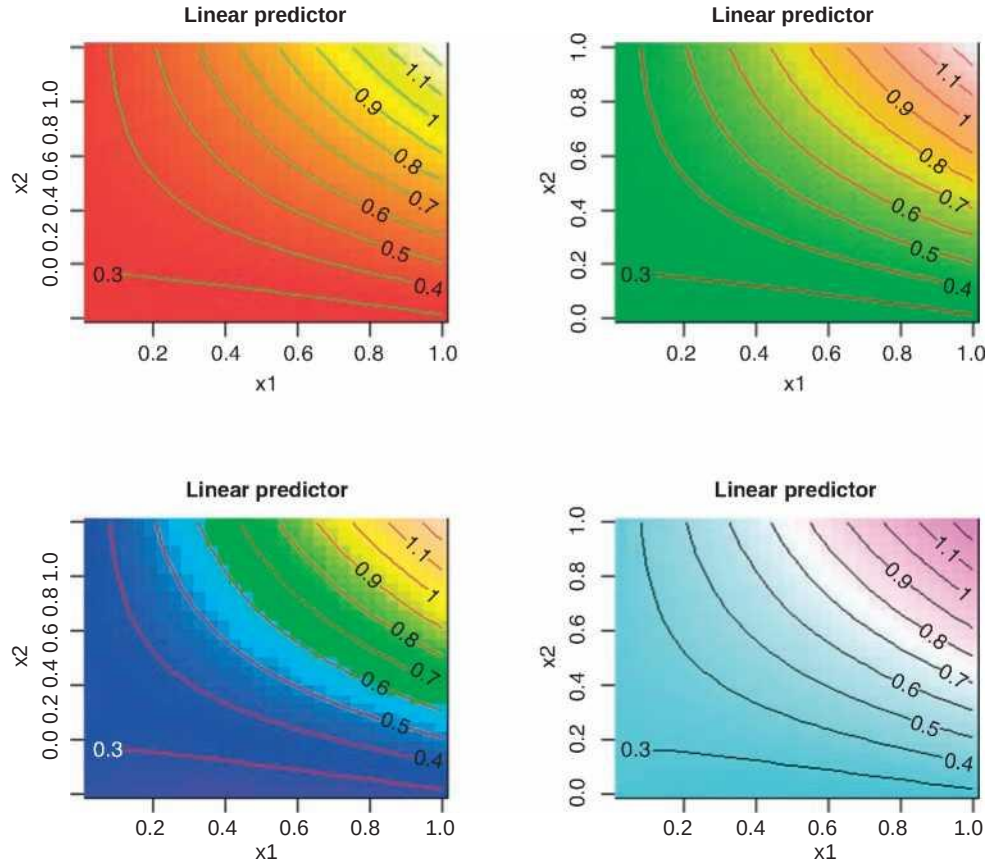


FIGURE 5.36 Contour plots of contour surfaces using vis.gam().

```
> b <- gam(y~fac+s(x))
> vis.gam(b,theta=-35,color="heat") # exemplu de factor
> # Ieșire: Figura 5.37 .
>
> z <- rnorm(80)*0,4
> y <- as.numeric(fac)+3*x^2*z+rnorm(80)*0.1
> b <- gam(y~fac+s(x,by=z))
> vis.gam(b,theta=-35,color="heat",cond=list(z=1))
> # „prin" exemplu de variabilă
> # Ieșire: Figura 5.38 .
>
> vis.gam(b,view=c("z","x"),theta= 35)
> # grafic împotriva după variabilă
> # Ieșire: Figura 5.39 .
```

(a) Descrieți, cu propriile cuvinte, semnificația și scopul fiecăreia dintre comenzile R din segmentul de cod precedent.
(b) R alternative care ar putea obține aceleași rezultate.
(c) Introduceți mediul R și rulați segmentul de cod pentru a obține Figura 5.39 .
(d) Urmând comanda > ls("pachet:mgcv") : Câte fișiere sunt listate sub pachetul mgcv ?
(e) mgcv este un program util. Urmând comanda > mgcv : Inspectați codul pentru această funcție.

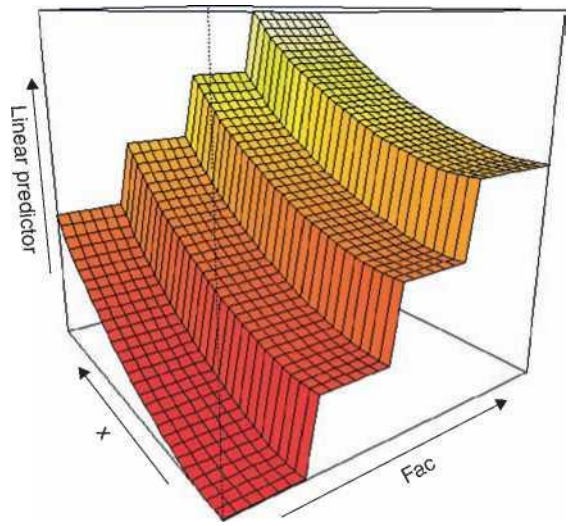


FIGURA 5.37 Grafice de contur ale suprafețelor folosind vis.gam() cu factor.

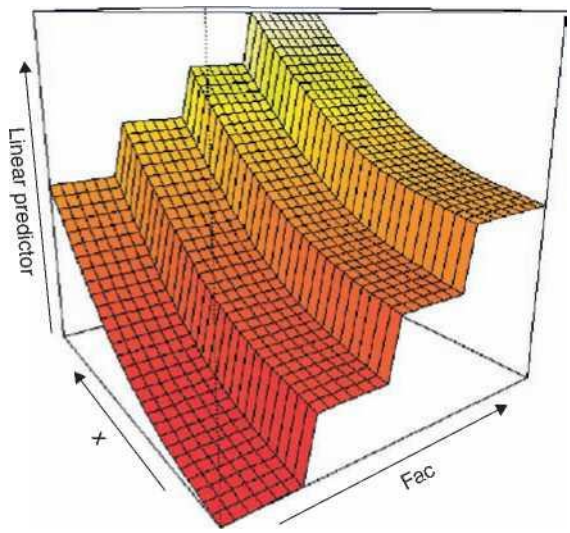


FIGURA 5.38 Grafice de contur ale suprafețelor folosind variabilele vis.gam() „by".

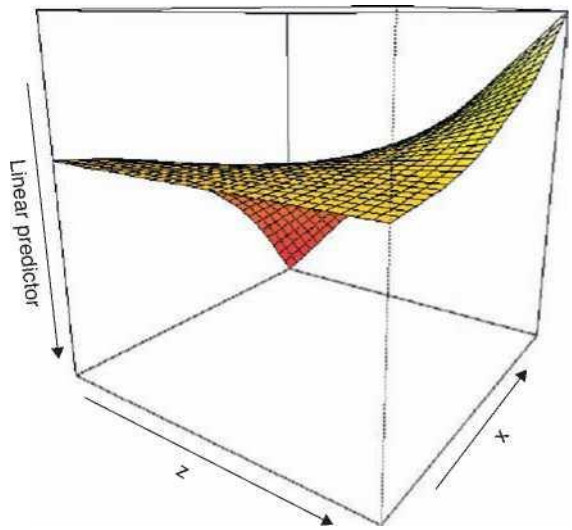


FIGURA 5.39 Grafice de contur ale suprafețelor folosind vis.gam() față de variabila „by”.

REFERINȚE

Birnbaum, ZW și Tingey, FH (1951). Contururi de încredere unilaterale pentru funcțiile de distribuție a probabilității. *Analele statisticii matematice, 22* (4), 592–596.

Conover, WJ (1971). *Statistici practice neparametrice* . New York, NY: John Wiley & Sons. pp. 295–301 (testul Kolmogorov cu o probă), 309–314 (testul Smirnov cu două probe).

Dalgaard, P. (2002). *Statistici introductive cu R* (Seria Statistică și Calcul). New York, NY: Springer.

Daniel, WW (2005). *Biostatistică: o bază pentru analiză în științele sănătății* (ediția a 7-a; Seria Wiley în Probabilitate și Statistică - Secțiunea Probabilitate și Statistică aplicată). New York, NY: John Wiley & Sons.

Durbin, J. (1973). *Teoria distribuției pentru teste bazate pe funcția de distribuție a eșantionului.* Philadelphia, PA: SIAM.

Hastie, T., & Tibshirani, R. (1990). *Modele aditive generalizate* .
Londra, Marea Britanie: Chapman & Hall.

Kaplan, EL, & Meier, P. (1958). Estimare neparametrică din observații incomplete. *Jurnalul Asociației Americane de Statistică, 53* , 457–481.

Kolmogorov, AN (1964). *Fundamentele teoriei probabilității.* New York, NY: Chelsea Publishing.

Marsaglia, G., Tsang, WW și Wang, J. (2003). Evaluarea distribuției lui Kolmogorov. *Journal of Statistical Software* , *8* (18), 1–4. Preluat de la http://www.jstatsoft.org/v08/i18/

Ng, KW și Tong, H. (2010, noiembrie). *Inversarea formulei Bayes și măsurile câștigului de informații bayesiene și dependenței pe perechi* (Raportul de cercetare nr. 477). Hong Kong, China: Universitatea din Hong Kong, Departamentul de Statistică și Știință Actuarială. Preluat de pe http://lx2.saas.hku.hk/research/research-report-477.pdf

Tan, MT, Tian, GL. și Ng, KW (2010). *Probleme bayesiene de date lipsă: EM, creșterea datelor și calcul neiterativ.* Boca Raton, FL: Chapman & Hall / CRC.

Triola, MM și Triola, MF (2006). *Biostatistica pentru stiintele biologice si ale sanatatii.* Boston, MA: Pearson/Addison Wesley.

Verzani, J. (2005). *Folosind R pentru statistici introductive.* Boca Raton, FL: Chapman & Hall/CRC.

Wood, SN (2006). *Modele de aditivi generalizate: O introducere cu R.* Boca Raton, FL: Chapman & Hall/CRC.

SIX

Studii de caz-control și studii de cohortă în epidemiologie

INTRODUCERE

[1] Cohort studies in epidemiology: http://en.wikipedia.org/wiki/Cohort_study

Studiile controlate randomizate (RCT) sunt considerate o metodologie superioară în ierarhia dovezilor în terapie, deoarece limitează potențialul oricăror prejudecăți prin atribuirea aleatorie a unui grup de caz-subiect unei intervenții și a unui alt grup de subiecți la neintervenție (sau un placebo). Acest lucru minimizează șansa ca incidența variabilelor de *confuzie* să difere între cele două grupuri. Studiile de cohortă pot fi efectuate fie prospectiv, fie retrospectiv. O examinare mai detaliată a RCT este prezentată în Capitolul 7 .

6.1 TEORIA ȘI ANALIZA STUDIILOR DE CAZ-CONTROL

Cercetările și investigațiile în epidemiologie și științele sănătății folosesc pe scară largă studiile caz-control. Astfel de studii pot identifica factorii care pot contribui la o afecțiune medicală prin compararea subiecților de caz care au acea boală/afecțiune (cazurile **)** cu subiecții care nu au boala/afecțiunea, dar sunt altfel similari (**martorii**). [2]

Un avantaj practic al studiului caz-control este că este relativ ieftin . De asemenea, acest tip de studiu poate fi (și este frecvent) întreprins de cercetători individuali, echipe mici sau facilități individuale.

Utilizarea studiului caz-control a condus la un număr considerabil de progrese științifice și descoperiri semnificative. În analele epidemiologiei, un succes remarcabil al acestui tip de studiu a fost demonstrarea relației dintre apariția cancerului pulmonar și utilizarea produselor din tutun de către Richard Doll și colab., care au demonstrat o asociere semnificativă statistic între cele două într-un studiu mare caz-control (Doll, Petro, Boreham și Sutherland, 2004). Deși oponenții subliniază în mod corect că acest tip de studiu nu poate dovedi prin el însuși cauzalitatea, eventualele rezultate ale studiilor de cohortă au confirmat legătura cauzală pe care o sugeraseră studiile caz-control: *fumatul este o cauză a aproximativ 87% din totalul mortalității prin cancer pulmonar în Statele Unite ale Americii.*

Avantajele și limitările studiilor de caz-control

Studiile caz-control tind să fie mai puțin costisitoare de realizat decât studiile de cohortă prospective și au potențialul de a fi mai scurte ca durată. Un alt avantaj este puterea statistică mai mare a acestui tip de studiu în mai multe situații; studiile de cohortă necesită adesea un număr suficient de evenimente de boală pentru a se acumula înainte de a putea oferi multe informații.

Cu toate acestea, studiile caz-control sunt de natură observațională și, prin urmare, nu oferă același nivel de dovezi ca RCT. Rezultatele pot fi confundate de diverși factori în măsura în care dau răspunsuri opuse celor ale unor studii mai bune. De asemenea, poate fi mai dificil să se stabilească cronologia expunerii la rezultatul bolii în cadrul unui studiu caz-control decât în cadrul unui proiect de studiu de cohortă prospectiv. În cel din urmă, expunerea este constatată înainte ca subiecții cazului să fie urmăriți în timp pentru a stabili starea lor de rezultat.

[2] Case-control studies in epidemiology:
http://en.wikipedia.org/wiki/Case-control_study

Cel mai important dezavantaj al studiilor caz-control se referă la dificultatea de a obține informații fiabile despre starea de expunere a unei persoane în timp. Studiile caz-control sunt, prin urmare, plasate jos în ierarhia dovezilor. Cu toate acestea, au fost efectuate multe studii caz-control de înaltă calitate și fiabile și au produs rezultate utile.

Analiza studiilor caz-control

Studiile caz-control au fost analizate inițial prin testarea dacă au existat diferențe semnificative între proporția de subiecți expuși între cazuri și controale . Dacă rezultatul bolii de interes este rar, raportul cotelor de expunere poate fi utilizat pentru a estima riscul relativ. Mai mult decât atât, raportul șanselor de expunere poate fi utilizat pentru a estima raportul ratei de incidență a expunerii în mod direct, fără a fi nevoie de ipoteza bolii rare.

Următorul este un exemplu de studiu caz-control în care calculele R au fost utilizate pentru a analiza seturile de date.

> **Exemplul 6.1:** Un studiu caz-control privind eficacitatea vaccinării BCG împotriva tuberculozei (TB)

Vaccinul Bacillus Calmette-Guerin (BCG) este utilizat pe scară largă împotriva tuberculozei (TB). Dezvoltat în anii 1930, este format dintr-o tulpină vie, slăbită de *Mycobacterium bovis* .

Colditz și colab. (1994) au raportat date din 13 studii clinice ale vaccinului BCG, fiecare investigând eficacitatea acestuia în tratamentul TBC. A fost înregistrat numărul de subiecți de caz care suferă de TBC cu sau fără vaccinare BCG. Setul de date conține, de asemenea, valorile altor două variabile pentru fiecare studiu: latitudinea geografică a locului în care a fost efectuat studiul și anul publicării. Aceste două variabile pot fi utilizate pentru a investiga orice eterogenitate între studii.

Sursa setului de date:
Colditz, GA, Brewer, TF, Berkey, CS, Wilson, ME, Burdick, E., Fineberg, HV și Mosteller, F. (1994). Eficacitatea vaccinului BCG în prevenirea tuberculozei *:* meta-analiză a literaturii publicate. *Journal of the American Medical Association, 271* (2), 698–702.

Următorul segment de cod R poate fi utilizat pentru a calcula un studiu caz-control pe baza setului de date disponibil:

```
> install.packages(„HSAUR")
> # Preluat din A Handbook of Statistical Analyses Using R (Everitt & Hothorn, 2006)
> bibliotecă (HSAUR)
Se încarcă pachetul necesar: zăbrele
Se încarcă pachetul necesar: MASS
Se încarcă pachetul necesar: scatterplot3d
Mesaje de avertizare:
1:  pachetul „HSAUR" a fost construit sub versiunea R 2.13.2
2:  pachetul „scatterplot3d" a fost construit sub versiunea R 2.13.2
> ls("pachet:HSAUR") # Notarea setului de date BCG în pachetul HSAUR
```

```
 [1] „agefat”              "aspirină"      "BCG"          "rata de naștere"
 [5] „cancer de            "BtheB"         " nori "              „CYGOB1”
 [9] „epilepsie”         „Forbes2000”      "foster"       "flori de gradina"
[13] „GHQ”               "heptatlon"       „HSAURtable”           "Lanza"
[17] „mastectomie”        "meteo"          "oraleiuni"    "fosfat"
[21] „grupuri de          "planete"        "plasma"              "polipi"
[25] „polipi3”            "ceramică"       "rearestari"   "respirator"
[29] „lățimea           "schizofrenie"      „schizofrenie2” „zile de școală”
[33] „cranii”             "fumat"          "studenti"             "sinucideri"
[37] „pastă de dinți”     "vot"            "apă"          "voluri de
[41] „valuri”           „creștere în       "rolul femeilor"
> date(BCG)
# Format set de date -
```

A data frame with 13 observations on the following 7 variables:

# Study	Un identificator al studiului.
# BCGTB	Numărul de subiecți care suferă de TBC *după* o vaccinare BCG.
# BCGVacc	Numărul de subiecți care au fost vaccinați cu BCG.
# NoVaccTB	Numărul de subiecți care suferă de TBC care *nu* au primit vaccinarea BCG.
# NoVacc	Numărul total de subiecți fără vaccinare BCG. Poziția geografică a
# Latitude	locului în care a fost efectuat studiul. Anul în care a fost întreprins
# Year	studiul.

> BCG *# Inspectarea setului de date*

	Studiu	BCGTB	BCGVacc	NoVaccTB	NoVacc	Latitudine	An
1	1	4	123	11	139	44	1948
2	2	6	306	29	303	55	1949
3	3	3	231	11	220	42	1960
4	4	62	13598	248	12867	52	1977
5	5	33	5069	47	5808	13	1973
6	6	180	1541	372	1451	44	1953
7	7	8	2545	10	629	19	1973
8	8	505	88391	499	88391	13	1980
9	9	29	7499	45	7277	27	1968
10	10	17	1716	65	1665	42	1961
11	11	186	50634	141	27338	18	1974
12	12	5	2498	3	2341	33	1969
13	13	27	16913	29	17854	33	1976

```
> atașează (BCG)
> boxplot(BCG$BCGTB/BCG$BCGVacc, # Două boxplot pe BCG
+ BCG$NoVaccTB/BCG$NoVacc,
+ nume = c("Vaccinare BCG", "Fără vaccinare BCG"),
+ ylab = „Procent de cazuri BCG")
> # Ieșire: Figura 6.1 .
```

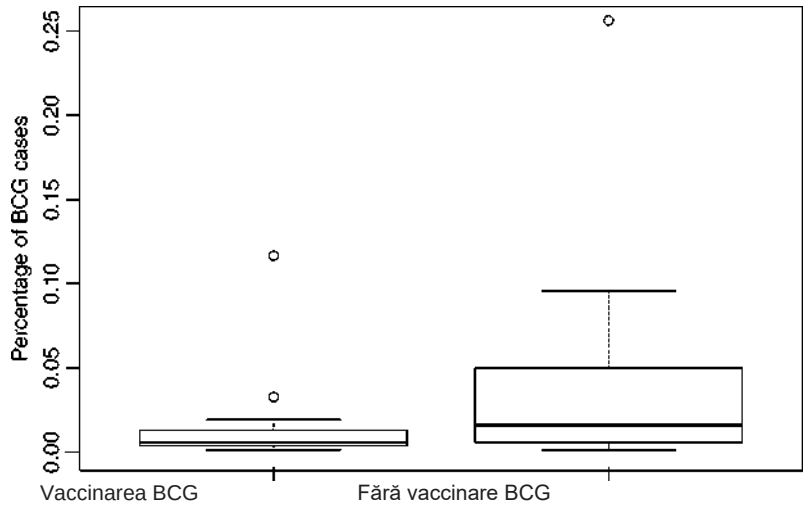


FIGURA 6.1 Boxplot pentru studiul caz-control privind eficacitatea vaccinării BCG în prevenirea TB.

Decizie biostatistică și concluzie: O comparație a celor două boxplot din Figura 6.1 arată clar că procentele de cazuri de TB pentru grupul fără vaccinare BCG sunt de câteva ori *mai mari* decât pentru grupul cu vaccinare BCG. Prin urmare, o concluzie preliminară rezonabilă este că populațiile care primesc vaccinare BCG vor avea *mai puține* șanse de a contracta TBC.

Observatii:

1. În cadrul de date BCG , informațiile de date au fost clasificate în funcție de șapte variabile: „Studiu" , „BCGTB" , „BCGVacc" , „NoVaccTB" , „NoVacc" , „Latitudine" și „An" .
2. Acest set de date se pretează la o analiză gata utilizând funcția boxplot() , diferențiind grupurile cu vaccinare BCG de grupurile fără vaccinarea BCG:
   ```
   > boxplot(BCG$BCGTB/BCG$BCGVacc,
   +           BCG$NoVaccTB/BCG$NoVacc, …
   ```
   Utilizarea acestei funcții a oferit imediat reprezentarea grafică a setului de date din Figura 6.1 .
3. Astfel, în analiza datelor din studiile caz-control, pregătirea cadrului de date este importantă pentru analiza directă a setului de date.
4. Alte analize de tip boxplot() pot fi efectuate cu astfel de seturi de date.

Întrebări de revizuire pentru Secțiunea 6.1

1. (a) Ce este un studiu caz-control?
   (b) Dați un exemplu de studiu caz-control în epidemiologie sau științele sănătății.
2. Care sunt principalele limitări ale unui studiu caz-control? Explicați și dați exemple.
3. (a) Pot fi întreprinse simultan un studiu caz-control și un studiu de cohortă în epidemiologie? Dacă da, cum? Care sunt avantajele acestei abordări?
   (b) Furnizați un exemplu al acestei abordări.

Exerciții pentru Secțiunea 6.1

1. În Exemplul 6.1 , care sunt funcțiile următorului segment de cod R utilizate în

analiza setului de date BCG ?

```
> boxplot(BCG$BCGTB/BCG$BCGVacc , # Două boxplot pe BCG
+          BCG$NoVaccTB/BCG$NoVacc,
+          nume = c("Vaccinare BCG", "Fără vaccinare BCG"),
+          ylab = „Procent de cazuri BCG")
```

2. Funcția boxplot() a fost utilizată în analiza eșantionului. Care sunt alte funcții R care ar putea produce rezultate similare?
3. Dacă cele două boxplot din Figura 6.1 ar fi reprezentate separat, ce segmente de cod R ați folosi? Demonstrați rezultatele.
4. În pachetul CRAN coxphf (regresia Cox cu probabilitatea penalizată a lui Firth), fișierul de date sân conține setul de date despre cancerul de sân utilizat de Heinze și Schemper (2001). Acest set de date conține informații despre 100 de pacienți cu cancer de sân, inclusiv timpul de supraviețuire, starea de supraviețuire, stadiul tumorii, statusul ganglionar, gradarea și expresia tumorii catepsinei-D.

   Descrieți ce realizează următorul segment de cod R pentru acest set de date:

```
>  date (sân)
> potrivire.sân <
+ coxphf(date=sân,Surv(TIME,CENS)~T+N+G+CD)
>  rezumat(fit.breast)
```

5. Executați segmentul de cod R din Exercițiul 3. Comentați rezultatele.
6. În pachetul de supraviețuire CRAN , vezica urinară a fișierului de date conține un set de date clinice cu informații despre 340 de subiecți de caz.
   (a) Descărcați acest set de date.
   (b) Descrieți ce realizează următorul segment de cod R pentru acest set de date:

```
# Potriviți un model stratificat, grupat pe pacienți
> vezică1 <- vezică[vezică$enum < 5, ]
> coxph(Surv(oprire, eveniment) ~ (rx + dimensiune + număr) *
+ strate(enum) + cluster(id), vezică1)
```

7. Executați segmentul de cod R din Exercițiul 5. Comentați rezultatele.

6.2 TEORIA ȘI ANALIZA STUDIILOR DE COORTE

O **cohortă** este un grup de subiecți de caz care împărtășesc o experiență sau o caracteristică comună (de exemplu, sunt expuși la un medicament, vaccin sau poluant, sau sunt supuși unei anumite proceduri medicale sau se nasc) într-o perioadă definită. Un grup de persoane care s-au născut într-o anumită zi sau într-o anumită perioadă formează o **cohortă de naștere** . Grupul de comparație poate fi populația generală din care este extrasă cohorta sau poate fi o altă cohortă de persoane despre care se crede că au avut o expunere mică sau deloc la substanța investigată, dar similară în rest. Alternativ, subgrupurile din cadrul cohortei pot fi comparate între ele.

Datele studiului de cohortă pot ajuta la determinarea factorilor de risc pentru contractarea de noi boli, deoarece un astfel de studiu este o observație longitudinală a individului în timp, în care datele sunt colectate la intervale regulate, reducând astfel eroarea de reamintire. O cohortă

studiul poate furniza dovezi care infirmă existența unei asocieri suspectate între cauză și

efect. În schimb, eșecul de a respinge o ipoteză întărește încrederea în acea ipoteză. Studiile de cohortă longitudinale prospective între boală și expunere ajută astfel la studiul asociațiilor cauzale, deși distingerea cauzalității adevărate necesită de obicei o coroborare suplimentară din alte surse.

Important este că *cohorta definită nu poate fi un grup de persoane care au deja boala* . Cohorta trebuie identificată înainte de apariția bolii care face obiectul anchetei. Ancheta urmărește, pentru o perioadă de timp, un grup de subiecți de caz care nu au boala și notează cine dezvoltă boala.

În general, studiile de cohortă sunt costisitoare de efectuat, sunt sensibile la uzură și necesită o urmărire extinsă pentru a genera date utile. Cu toate acestea, calitatea rezultatelor din studiile de cohortă pe termen lung este substanțial superioară celor obținute din studii retrospective sau transversale. *Studiile de cohortă prospective oferă cele mai fiabile rezultate în epidemiologia observațională.* Acestea permit studierea unei game largi de asocieri expunere-boală.

De exemplu, unele studii de cohortă urmăresc grupuri de copii de la naștere și înregistrează o gamă largă de informații despre aceștia. Valoarea unui studiu de cohortă depinde adesea de capacitatea cercetătorilor de a rămâne în contact cu toți membrii cohortei; unele studii au continuat decenii.

O aplicație importantă a studiilor de cohortă

Un exemplu de întrebare epidemiologică la care se poate răspunde prin utilizarea unui studiu de cohortă este: **Expunerea la *X* (să zicem, fumatul) se asociază cu rezultatul *Y* (să zicem, cancerul pulmonar)?**

Un astfel de studiu ar recruta un grup de fumători și un grup de nefumători (grupul neexpus), îi va urmări pentru o perioadă stabilită de timp și ar observa diferențe în incidența cancerului pulmonar între grupuri la sfârșitul acestui timp. Grupurile sunt *corelate* în ceea ce privește multe alte variabile, cum ar fi starea economică și alte stări de sănătate, astfel încât variabila care este evaluată, **variabila independentă** (în acest caz, fumatul) poate fi izolată ca cauză a **variabilei dependente** (în acest caz). caz, cancer pulmonar). În acest exemplu, o creștere **semnificativă statistic** a incidenței cancerului pulmonar în grupul de fumători în comparație cu grupul de nefumători este o dovadă în favoarea ipotezei. Cu toate acestea, rezultatele rare, cum ar fi cancerul pulmonar, nu sunt, în general, studiate prin utilizarea cohortelor; în schimb, se folosesc studii caz-control.

Studii clinice

O mare parte din cercetarea medicală realizează studii **clinice pe termen mai scurt** . Astfel de studii urmăresc în mod obișnuit două grupuri de pacienți pentru o perioadă de timp și compară un punct final sau o măsură de rezultat între cele două grupuri.

Studii controlate randomizate

Un RCT este o metodologie superioară în ierarhia dovezilor, deoarece limitează potențialul de părtinire prin atribuirea aleatorie a unui grup de caz-subiect unei intervenții.

și un alt grup de caz-subiect la un placebo, minimizând astfel șansa ca incidența

variabilelor de confuzie să difere între cele două grupuri. Cu toate acestea, uneori nu este etic sau practic să se efectueze un RCT pentru a rezolva o întrebare clinică . De exemplu, dacă cineva are deja dovezi rezonabile că fumatul cauzează cancer pulmonar, a convinge un grup de nefumători să se apuce de fumat pentru a testa această ipoteză ar fi, în general, considerat neetic.

Studii de cohortă pentru bolile la alegere și bolile netransmisibile (Handysides & Landless, 2012)

În 2011, Secretarul General al Națiunilor Unite, abordând crizele globale cauzate de creșterea rapidă a **bolilor netransmisibile (BNT)** – care sunt în mare parte boli prevenibile – a declarat: „Colaborarea noastră este mai mult decât o necesitate de sănătate publică. BNT reprezintă o amenințare la adresa dezvoltării. NCD-urile îi lovesc în mod deosebit pe cei săraci și vulnerabili și îi conduc mai adânc în sărăcie" (Handysides & Landless, 2012) . Aspectul exterior a fost sumbru din cauza incidenței în creștere rapidă a BNT la nivel mondial, țările mai sărace și emergente confruntându-se cu cele mai mari provocări.

La nivel global, deși eforturile continue sunt îndreptate către bolile transmisibile și infecțioase (cum ar fi malaria, TBC, gastroenterita, SIDA, HIV etc.), BNT cresc rapid. Ele sunt atât o cauză majoră a deceselor care pot fi prevenite, cât și o contribuție majoră la pierderea productivității și la sărăcie. BNT includ:

- Boli de inima
- Accident vascular cerebral
- Cancer
- Boli respiratorii cronice
- Diabet

Acestea afectează toate comunitățile și oamenii. Principalii lor factori de risc sunt similari la nivel mondial și sunt bine cunoscuți, inclusiv utilizarea și efectele:

- Tutun, sub toate formele sale (inclusiv fumatul pasiv)
- Alcoolul, sub toate formele sale
- Exces de sare și zahăr în dietă
- Alimente bogate în grăsimi trans și saturate
- Obezitatea
- Inactivitate fizică

Din punctul de vedere al biostatisticii, este clar că stabilirea relației cauză-efect dintre acești factori de risc și BNT poate fi abordată folosind studii de cohortă. Un astfel de studiu ar compara toate cohortele care s-au implicat în stiluri de viață care includ unul sau mai mulți dintre principalii factori de risc enumerați anterior și ar compara incidența lor respectivă de BNT.

CATEVA STUDII DE COORTA BINE CUNOSCUTE

Două studii de cohortă care se desfășoară de mai bine de 50 de ani sunt Fram ingham Heart Study și National Child Development Study (raportat în

Jurnalul *Internațional de Epidemiologie*). Ultimul studiu compară două cohorte:

„Studiul Millennium Cohort" (Statele Unite) și „King's Cohort" (Regatul Unit).

Cel mai mare studiu de cohortă la femei este Studiul de sănătate a asistentelor. Începută în 1976, urmărește peste 120.000 de asistente și a fost analizat pentru multe afecțiuni și rezultate diferite.

Cel mai mare studiu de cohortă din Africa este studiul *Birth to Twenty* , care a început în 1990 și urmărește o cohortă de peste 3.000 de copii născuți în săptămânile de după eliberarea lui Nelson Mandela din închisoare.

Studiile Adventiste de Sănătate (AHS-1 și AHS-2) au înrolat între 30.000 și 100.000 de subiecți de caz.

În toate aceste investigații, subiecții expuși și neexpuși sunt observați până când dezvoltă un rezultat de interes. O abordare analitică perspicace este utilizarea instrumentelor de analiză a supraviețuirii (de exemplu, procedura Kaplan–Meier [K–M], modelul de regresie Cox sau diagrame de hazard proporțional), toate acestea putând fi executate în mediul R (vezi Capitolul 5). Următoarele sunt câteva exemple tipice.

■ **Exemplul 6.2:** Studii de cohortă utilizând modelul de regresie Cox și procedura KM - pentru a analiza un set de date privind fumul de tutun pasiv

Tammemagi, Neslind-Dudas, Simoff și Kvale (2004) au raportat despre rolul comor bidității și al tratamentului pentru fumat și supraviețuirea cancerului pulmonar a 1155 de subiecți de caz. Eșantionul a fost format din 470 de femei (41%), 685 de bărbați (59%), 462 de negri (40%) și 693 de albi (60%). Obiectivul a fost de a determina dacă fumatul prezice în mod independent supraviețuirea la pacienții cu cancer pulmonar sau dacă un efect existent este mediat de comorbiditate și/sau tratament.

Abordarea studiului:
Analiza riscurilor proporționale Cox a fost utilizată pentru a studia cohorta de 1.155 de pacienți cu cancer pulmonar diagnosticați la Henry Ford Health System între 1995 și 1998, inclusiv.

Rezultate:

- Ajustat pentru covariatele inițiale de vârstă, sex, consum de droguri ilicite, simptome - adverse , histologie și stadiu, raportul de risc (HR) pentru fumat (actual vs. anterior/niciodată) a fost 1,37 (IC 95% [1,18, 1,59] ; $p < .001$).
- Ajustat pentru covariatele inițiale și pentru 18 comorbidități dăunătoare, HR pentru fumat a fost 1,38 (IC 95% [1,18, 1,60]; $p < 0,001$), indicând că efectul periculos al fumatului nu a fost mediat de comorbiditate. Fumatul curent a fost invers asociat cu tratamentul (orice intervenție chirurgicală și/sau chimioterapie și/sau radioterapie vs. niciuna; odds ratio, 0,73; IC 95% [0,55, 0,98]; $p = 0,03$).
- Ajustat pentru covariabile de bază , comorbidități și tratament, HR pentru fumătorul actual versus fostul/niciodată a fost de 1,26 (IC 95% [1,08, 1,47]; $p = 0,003$), o scădere de 30,7% care a fost explicată prin tratament (HR pentru orice tratament vs. niciunul: 0,40 IC 95% [0,33, 0,48] *;* < .001).

Supraviețuirea mediană pentru fumătorii actuali a fost de 0,76 ani (IC 95% [0,67, 0,89]), iar pentru foștii fumători/niciodată a fost de 1,01 ani (IC 95% [0,89, 1,15]). Vezi Figura 6.2 .

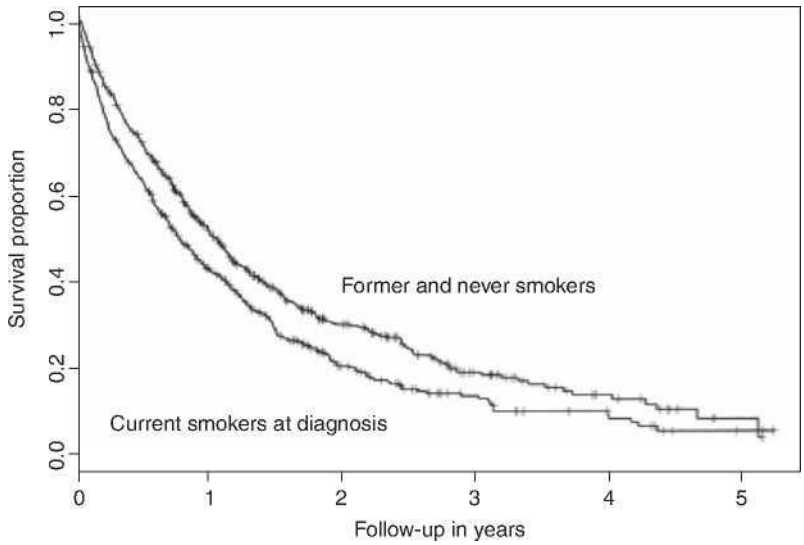


FIGURA 6.2 Graficul de supraviețuire K–M pentru 1155 de subiecți cu cancer pulmonar, stratificat în funcție de statutul de fumat.

Concluzii:
Fumatul curent la diagnosticare este un predictor independent important al supraviețuirii scurte a cancerului pulmonar. Faptul că acest efect nu a fost explicat de factori socio-demografici și de expunere, simptome adverse, histologie, stadiu, comorbiditate și tratament sugerează că poate fi mediat prin efecte biologice directe.

- **Exemplul 6.3:** Studiu de cohortă folosind modelul de regresie proporțională Cox

În pachetul CRAN coxrobust , funcția coxr() este disponibilă pentru ajustarea eficientă și robustă a modelului de regresie a hazardelor proporționale Cox în forma sa de bază, unde variabilele explicative sunt independente de timp, cu un eveniment per subiect. Metoda se bazează pe o modificare lină a probabilității parțiale.

Abordarea este de a maximiza o funcție obiectivă care este o modificare lină a probabilității parțiale. Observatii cu valori excesive de

$$L(T)\exp(b' Z)$$

unde:
L = pericolul cumulat,
b = vectorul parametrilor și b' = transpunerea lui b
Z = unele variabile explicative și
T = timpul de supraviețuire posibil cenzurat,
sunt ponderate în jos. Atât L, cât și b sunt estimate robust iterativ.

Rezultatele numerice sunt susținute de funcția plot() , un instrument grafic, care într-o serie de cinci grafice compară cât de bine sunt corelate datele de modelul de riscuri proporționale estimate cu metode non-robuste (în negru) și robuste (în verde):

- Primul grafic arată diferența standardizată a două funcții de supraviețuire estimate , una folosind modelul Cox și cealaltă folosind estimatorul K–M.

- Celelalte patru grafice arată aceleași diferențe pentru patru straturi, definite de quartilele predictorului liniar estimat.

Compararea rezultatelor estimărilor, împreună cu analiza graficelor, poate oferi informații foarte detaliate despre potrivirea modelului.

Utilizare:

coxr(formula, date, subset, na.action, trunc = 0,95,
f.greutate = c(„liniară", „pătratică", „exponențială"),
singular.ok = TRUE, model = FALSE)

Argumente:

formula Un obiect formulă, cu răspunsul în stânga unui operator ~ și termenii în dreapta. Răspunsul trebuie să fie un obiect de supraviețuire, așa cum este returnat de funcția Surv() .

date Un cadru de date pentru interpretarea variabilelor numite în formulă sau în subset .

subset Expresie care spune că numai un subset al rândurilor de date ar trebui utilizat în potrivire.

na.action O funcție de filtrare a datelor lipsă, aplicată modelului.cadru după ce a fost folosit orice argument subset .

trunc Aproximativ, o cuantilă a probei $T_i \exp(b' Z_i)$; determină nivelul de tăiere pentru estimatorul robust.

f.greutate Tipul funcției de ponderare; implicit este „quadratic" .

singular.ok valoare logică care indică modul de gestionare a coliniarității în matricea modelului. Dacă TRUE , programul va sări automat peste coloanele matricei X care sunt combinații liniare ale coloanelor anterioare. În acest caz, coeficienții pentru astfel de coloane vor fi NA , iar matricea de varianță va conține zerouri. Pentru calculele auxiliare, cum ar fi predictorul liniar, coeficienții lipsă sunt tratați ca zerouri.

model O valoare logică care indică dacă cadrul modelului ar trebui inclus ca o componentă a valorii returnate.

Următorul segment de cod R poate fi utilizat pentru a calcula un studiu de cohortă pe baza a două seturi de date disponibile:

1. plămân - datele despre cancerul pulmonar de la Mayo Clinic
2. veteran - date despre cancerul pulmonar al Administrației Veteranilor

```
> # Folosind plămân, datele despre cancerul pulmonar de la Mayo Clinic
> install.packages ("coxrobust")
> bibliotecă (coxrobust)
```

Se încarcă pachetul necesar: supraviețuire
Se încarcă pachetul necesar: spline

```
> ls("pachet:coxrobust")
```

> 1] „coxr" „gen_data" „plot.coxr" „print.coxr"

```
> date (plămân)
```

> ataşează (plămân)
> plămân

	inst	timp	starea	vârstă	Sex	ph.ecog	ph.karno	pat. karno	masă. cal	pierdere în
1	3	306	2	74	1	1	90	100	1175	N / A
2	3	455	2	68	1	0	90	90	1225	15
3	3	1010	1	56	1	0	90	90	N / A	15
4	5	210	2	57	1	1	90	60	1150	11
5	1	883	2	60	1	0	100	90	N / A	0
224	1	188	1	77	1	1	80	60	N / A	3
225	13	191	1	39	1	0	90	90	2350	-5
226	32	105	1	75	2	2	60	70	1025	5
227	6	174	1	66	1	1	90	100	1075	1
228	22	177	1	58	2	1	80	90	1060	0

```
> #folosiți datele despre cancerul pulmonar de la Mayo Clinic pentru
> #comparați rezultatele estimărilor nesolide și robuste
> rezultat <- coxr(Surv(time, status) ~ varsta + sex + ph.karno +
>                  meal.cal + wt.loss, data = plămân)
> rezultat # Ieşire:
```

Apel:
coxr(formula = Surv(timp, status) ~ vârstă + sex + ph.karno + meal.cal + wt.loss, date = plămân)

Estimator de probabilitate parțială

	coef	exp(coef)	se(coef)	p
vârstă	1.25e-02	1.013	0,011686	0,2844
sex	-4,73e-01	0,623	0,197557	0,0166
ph.karno	-9.64e-03	0,990	0,0071260	1764
masa.cal	-8,96e-05	1.000	0,000245	0,7146
pierdere în	-2.82e-03	0,997	0,006989	0,6868

testul Wald= 11.6 pe 5 df, p 0,0408

Estimator robust

	coef	exp(coef)	se(coef)	p
vârstă	0,004927	1.005	0,013755	0,720210
sex	-0,839739	0,432	0,261861	0,001342
ph.karno	-0,033673	0,967	0,009986	0,000746
masa.cal	-0,000397	1.000	0,000336	0,236812
pierdere	-0,008365	0,992	0,009398	0,373400

Testul Wald extins=23,3 pe 5 df, p=0,000292

```
> complot (rezultat)
```

Se așteaptă confirmarea modificării paginii... *# Ieșire* Figura 6.3 .
>
> *# Ieșire:* Figura 6.4 .
>
> *Folosind datele despre cancerul pulmonar de la* ***Administrația*** Veteranilor
> date (veteran)
> atașați (veteran)
Următoarele obiecte sunt mascate din „plămân": vârstă, stare, timp
> veteran

	trt	tipul	timp	starea	karno	diagtime	vârstă	anteri
1	1	scuamoase	72	1	60	7	69	0
2	1	scuamoase	411	1	70	5	64	10
3	1	scuamoase	228	1	60	3	38	0
4	1	scuamoase	126	1	60	9	63	10
5	1	scuamoase	118	1	70	11	65	10
133	2	mare	133	1	75	1	65	0
134	2	mare	111	1	60	5	64	0
135	2	mare	231	1	70	18	67	10
136	2	mare	378	1	80	4	65	0
137	2	mare	49	1	30	3	37	0

> *# Utilizați datele privind cancerul pulmonar al administrației veteranilor*
> *# pentru a compara rezultatele estimărilor nesolide și robuste*
> rezultat <- coxr(Surv(time,status) ~ age + trt + celltype + karno
> diagtime + anterior, date = veteran)
> rezultat
Apel: coxr(formula = Surv(time, status) ~ age + trt + celltype + karno + diagtime + prior, data = veteran)
Estimator de probabilitate parțială

	coef	exp(coef)	se(coef)	p
vârstă	-0,008739	0,991	0,00931	3.48e-01
trt	0,289208	1.335	0,20757	1.64e-01
celltypesmallcell	0,855501	2.353	0,27517	1.88e-03
tipul celularadeno	1,188162	3.281	0,30092	7.87e-05
tipul celulelor mari	0,392164	1.480	0,28260	1.65e-01
karno	-0,032808	0,968	0,00551	2.64e-09
diagtime	0,000191	1.000	0,00913	9.83e-01
anterior	0,007261	1.007	0,02323	7.55e-01

Testul Wald=62,2 pe 8 df, p=1,71e-10 Estimator robust

	coef	exp(coef)	se(coef)	p
Vârstă	**-0,01184**	**0,988**	**0,01336**	**3.75e-01**
Trt	**0.20132**	**1.223**	**0,24286**	**4.07e-01**
Tipuri de	**1,14056**	**3.129**	**0,43445**	**8.66e-03**
Celltypeadeno	**1,22137**	**3.392**	**0,49300**	**1.32e-02**
Tipul celular mare	**0,20405**	**1.226**	**0,46553**	**6.61e-01**
Karno	**-0,04230**	**0,959**	**0,00687**	**7.26e-10**
Diagtime	**-0,00489**	**0,995**	**0,00981**	**6.18e-01**
înainte	**0,01335**	**1.013**	**0,03447**	**6.98e-01**

Testul Wald extins=62 pe 8 df, p=1,9e-10

> complot (rezultat)

Se așteaptă confirmarea modificării paginii. # *Ieșirea* Figura 6.5 .

> # *Ieșire:* Figura 6.6 .

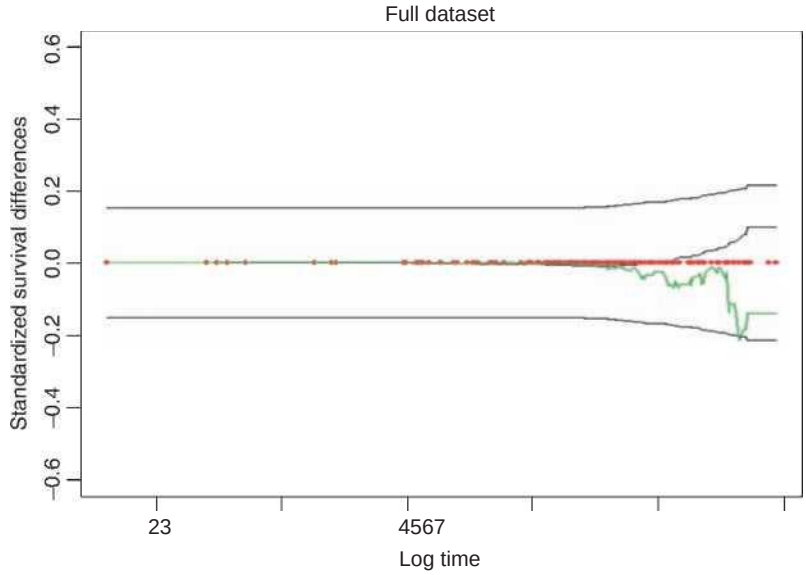


FIGURE 6.3 Cohort study using lung (the lung cancer data at Mayo Clinic). Standardized difference of two estimated survival functions, one via the Cox model and the other via the K–M estimator: nonrobust (black) and robust (green).

Studiile de cohortă și diagrama lexicului [24]în Biostatistica demografiei

Biostatistica demografică care se ocupă de studiul populațiilor umane utilizează adesea o **diagramă Lexis** (numită după economistul Wilhelm Lexis). Această diagramă bidimensională reprezintă evenimente (cum ar fi decesele sau nașterile) care se întâmplă

[24] Diagrama Lexis: http://en.wikipedia.org/wiki/Lexis_diagram

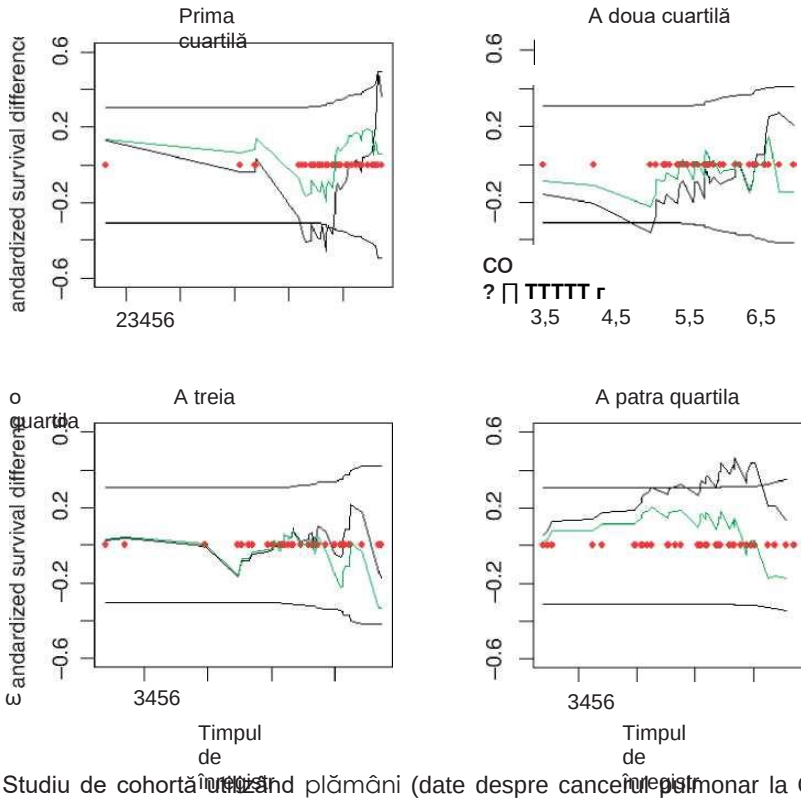


FIGURA 6.4 Studiu de cohortă utilizând plămâni (date despre cancerul pulmonar la Clinica Mayo). Diferența standardizată a două funcții de supraviețuire estimate: una prin modelul Cox și cealaltă prin estimatorul K–M . Cele patru grafice arată aceleași diferențe pentru patru straturi, definite de quartilele predictorului liniar estimat: nonrobust (negru) și robust (verde).

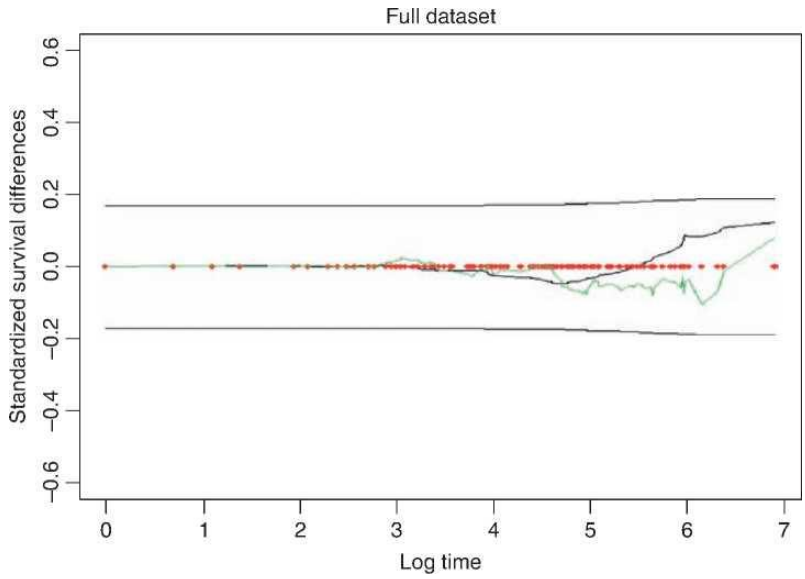


FIGURA 6.5 Studiu de cohortă utilizând veteran (date privind cancerul pulmonar de la Veterans Administration). Diferența standardizată a două funcții de supraviețuire estimate: una prin modelul Cox și cealaltă prin estimatorul K–M : nerobustă (negru) și robustă (verde).

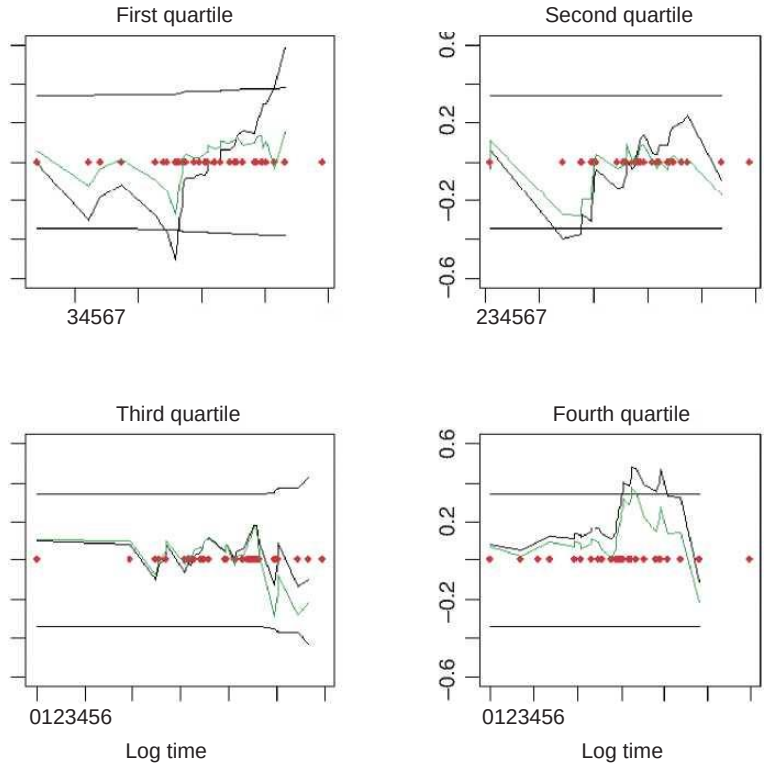


FIGURE 6.6 Cohort study using veteran (Veterans Administration lung cancer data). Standardized difference of two estimated survival functions: one via the Cox model and the other via the K–M estimator. The four graphs show the same differences for four strata, defined by the quartiles of the estimated linear predictor: nonrobust (black) and robust (green).

indivizi aparținând unor cohorte diferite. Ora calendaristică este de obicei reprezentată pe axa orizontală, iar vârsta este reprezentată pe axa verticală. De exemplu, decesul unui individ în 1988 la vârsta de 83 de ani este reprezentat de punctul (1988, 83); cohorta tuturor persoanelor născute în 1937 este reprezentată de o linie diagonală care începe de la (1937, 0) și continuă prin (1938, 1), (1939, 2), (1940, 3), …, (2000, 63), …, (2012, 75), …, (2037, 100) și așa mai departe.

Un exemplu de diagramă Lexis tipică este prezentat în Figura 6.7 .

Întrebări de revizuire pentru Secțiunea 6.2

1. (a) În investigațiile epidemiologice, ce este o cohortă? Da un exemplu.
   (b) Ce sunt studiile de cohortă, așa cum sunt utilizate în cercetări și investigații în epidemiologie și științele sănătății? Dați un exemplu de studiu de cohortă tipic.
2. (a) „Cohorta nu poate fi definită ca un grup de persoane care au deja boala.” Explica.
   (b) „Un studiu de cohortă poate fi o sarcină consumatoare de timp și poate fi foarte costisitor de realizat.” Explica.

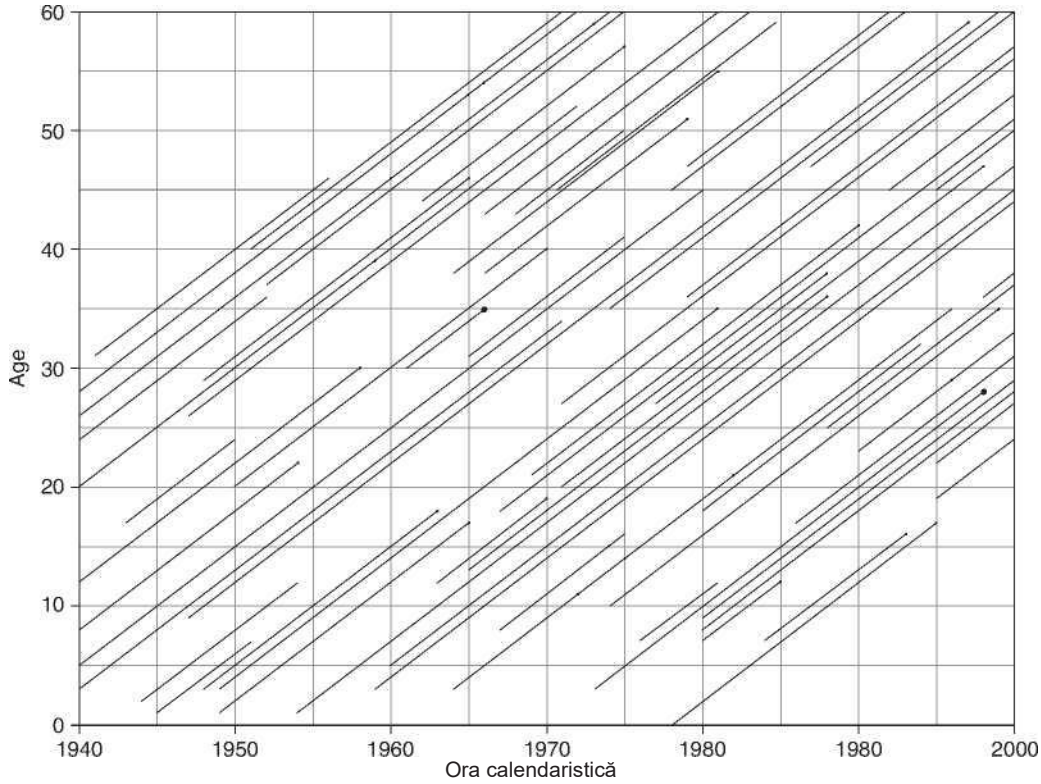


FIGURA 6.7 O diagramă Lexis tipică pentru biostatistica demografică.

3. (a) „Studiile de cohortă sunt potrivite pentru investigații în bolile la alegere și bolile netransmisibile." Explica.
   (b) Numiți cinci boli care sunt adecvate pentru investigare prin studii de cohortă.
4. (a) Numiți cinci exemple binecunoscute de studii de cohortă, fiecare implicând mii de subiecți de caz.
   (b) Numiți două abordări în analiza supraviețuirii care sunt potrivite pentru studiile de cohortă.
5. (a) Ce este o diagramă Lexis?
   (b) „O diagramă Lexis poate fi potrivită pentru analiză în studiile de cohortă." Explica.

Exerciții pentru Secțiunea 6.2

1. În Exemplul 6.3, care sunt funcțiile următorului segment de cod R utilizat în analiza plămânului setului de date ?
   ```
   > rezultat <- coxr(Surv(time, status) ~ varsta + sex + ph.karno +
   +                  meal.cal + wt.loss, date = plămân)
   ```
2. În loc de funcția coxr() utilizată în analiză, care sunt alte funcții R care pot produce rezultate similare?
3. În Exemplul 6.3, care sunt funcțiile următorului segment de cod R utilizate în analiza veteranului setului de date ?
   ```
   > rezultat <- coxr(Surv(time,status) ~ age + trt + celltype + karno
   ```

```
+                    diagtime + anterior, date = veteran)
```

4. În loc de funcția coxr() utilizată în analiza de la întrebarea 3, care sunt alte funcții R care pot produce rezultate similare?
5. În pachetul CRAN survival , funcția survfit() calculează funcția de supraviețuire prezisă pentru un model de hazarduri proporționale Cox. Forma generală de utilizare este
```
> survfit(formula, newdata,
+ se.fit=TRUE, conf.int=.95,
+       individ=FALSE,
+       tip,vartype,
+       conf.type=c("log","log-log","plain","none"),
+       censor=TRUE, id, ...)
```
Care este semnificația fiecărui argument din funcția survfit() ? (***SUGESTIE*** : Accesați site-ul web CRAN și găsiți pagina survfit.coxph{survival} pentru definițiile argumentelor acestei funcții.)
6. În cadrul pachetului de supraviețuire CRAN , următorul segment de cod R se potrivește unui model K–M pentru setul de date aml și apoi trasează curba K–M:
```
> fit <- survfit(Surv(time, status) ~ x, data = aml)
> plot(potrivire, lty = 2:3)
> legendă(100, .8, c(„Menținut", „Neîntreținut"), lty = 2:3)
```
(a) Care sunt funcțiile acestui segment de cod R utilizate în analiza setului de date aml ?
(b) Rulați segmentul de cod precedent într-un mediu R.
(c) Descrie rezultatele.
7. În cadrul pachetului de supraviețuire CRAN , următorul segment de cod R se potrivește unui model de riscuri proporționale Cox pentru setul de date ovarian (pentru un subiect de caz de 60 de ani), apoi trasează curba modelului:
```
> fit <- coxph(Surv(futime, fustat) ~ age, data = ovarian)
> plot(survfit(fit, newdata=data.frame(age=60)),
+       xscale=365,25, xlab = „Ani", ylab="Survival")
```
(a) Care sunt funcțiile acestui segment de cod R utilizate în analiza setului de date ovarian ?
(b) Rulați segmentul de cod precedent într-un mediu R.
(c) Descrie rezultatele.
8. În cadrul pachetului de supraviețuire CRAN , următorul segment de cod R se potrivește unui model K–M (pentru timpul până la progresie/deces pentru pacienții cu gammapatie monoclonală), pentru setul de date mgus1 și apoi trasează curbele de risc concurente ale modelului pentru incidența cumulativă:
```
> fit1 <- survfit(Surv(stop, event=='progresie') ~1,
+                 date=mgus1, subset=(start==0))
> fit2 <- survfit(Surv(stop, status) ~1, data=mgus1,
+                 subset=(start==0), etype=eveniment)
> # Riscuri concurente:
> # Curbele CI sunt reprezentate de la 0 în sus, mai degrabă decât
> # de la 1 în jos
> plot(fit2, fun='eveniment', xscale=365.25, xmax=7300,
```

```
+        mark.time=FALSE,
+        col=2:3, xlab="Ani după diagnosticarea MGUS")
> linii(fit1, fun='eveniment', xscale=365.25, xmax=7300,
+ mark.time=FALSE, conf.int=FALSE)
> text(10, .4, „Risc de concurență: moarte", col=3)
> text(16, .15,"Risc de concurență: progresie", col=2)
> text(15, .30,"KM:prog")
```

(a) Care sunt funcțiile acestui segment de cod R utilizate în analiza setului de date mgus1 ?

(b) Rulați segmentul de cod precedent într-un mediu R.

(c) Descrie rezultatele.

9. Pachetul CRAN survrec .

Un complex mioelectric migrator [sau complex motor migrator (MMC)] este un val de activitate bioelectrică care trece prin intestine într-un ciclu regulat în timpul postului. Aceste complexe ajută la declanșarea undelor peristaltice, care facilitează mișcarea substanțelor nedigerabile (fibre și corpi străini) din stomac, prin intestinul subțire, dincolo de sfincterul ileocecal și în colon. MMC își are originea în stomac, aproximativ la fiecare 80 de minute între mese și este responsabil pentru zgomotul stomacal experimentat atunci când este foame. MMC durează aproximativ 15 minute. De asemenea, servește la transportul bacteriilor din intestinul subțire la intestinul gros și la inhibarea migrației bacteriilor colonice în ileonul terminal. MMC poate fi parțial reglat de motilină; este inițiat în stomac ca răspuns la stimularea vagală şi nu depinde direct de nervii extrinseci.

În pachetul CRAN survrec , funcția mlefrailty.fit() este un estimator al funcției de supraviețuire pentru datele de timp de recurență corelate în cadrul unui model de fragilitate gamma care utilizează criteriul de probabilitate maximă. Obiectul rezultat al clasei survfitr poate fi reprezentat de funcția plot.survfitr() înainte de a fi returnat.

Forma de utilizare a acestei funcții mlefrailty.fit() este mlefrailty.fit(x,tvals, lambda=NULL, alpha=NULL, alpha.min, alpha.max, tol=1e-07, maxiter=500,alpha.console= ADEVĂRAT)

pentru care argumentele sunt:

x — Un obiect de eveniment recurent de supraviețuire.

tvals — Vector de timpi în care funcția de supraviețuire poate fi estimată.

lambda — Vector opțional al probabilităților de pericol de bază la t (a se vedea detalii în pachet survrec). Implicit este numdeaths/apply (AtRisk, 2,-sum).

alfa — Parametru opțional de formă și scară pentru distribuția de fragilitate bution. Dacă acest parametru este necunoscut, este estimat printr-un algoritm de maximizare așteptare (EM). Se calculează o sămânță pentru a obține convergența acestui algoritm (vezi detalii în pachetul survrec).

alpha.min — Limită stângă opțională a parametrului alfa; folosit pentru a obține o sămânță pentru a estima parametrul alfa. Valoarea implicită este 0,5.

alpha.max Limitare la dreapta opțională a parametrului alfa; folosit pentru a obține o sămânță pentru a estima parametrul alfa. Valoarea implicită este numărul maxim de momente distincte ale evenimentelor.

tol	Toleranța opțională a algoritmului EM utilizat pentru estimarea parametrului alfa. Implicit este 10e - 7.
maxiter	Numărul maxim opțional de iterații ale algoritmului EM utilizat pentru a estima parametrul alfa. Implicit este 500.
alpha.console	Dacă TRUE se tipărește în consolă, programul estimează valoarea inițială pentru alfa și estimarea alfa prin algoritmul EM; dacă este FALS , nu este.

Observatii:

1. O alegere comună a distribuției de fragilitate este o distribuție gamma cu parametrii de formă și scară setați egali cu un parametru necunoscut α . Funcția de supraviețuire marginală comună poate fi exprimată ca

$$F(t) = [\alpha / (\alpha + \Lambda_0(t))]^{\alpha} \quad (6.1)$$

Parametrul α controlează gradul de asociere între timpii de interapariție într-o unitate. Se poate demonstra că estimarea α și α_0 poate fi obținută prin maximizarea funcției de probabilitate marginală și a algoritmului EM.

Pentru a obține o convergență bună, mai întâi se estimează α . Această estimare este utilizată ca valoare inițială în procedura EM și este realizată prin maximizarea probabilității profilului pentru α . În acest caz, argumentele funcției mlefrailty. fit() , numit alpha.min și alpha.max , sunt limitele acestei maximizări. Maximul se obține folosind metoda de căutare a secțiunii de aur.

2. *Valoare:* Dacă nu se obține convergența algoritmului EM, valoarea inițială a lui α poate fi folosită ca argument alpha.min și recalculată.

n	Numărul de unități sau subiecte observate.
m	Vector al numărului de recurențe la fiecare subiect (lungimea n).
a eșuat	Vector al numărului de recurențe la fiecare subiect (lungime n*m). Vector ordonat (de exemplu, timpi ai primei unități, ori ai celei de-a doua unități, ..., ori ai n-unității).
cenzurat	Vector de timpi de cenzură pentru fiecare subiect (lungimea n).
numdistinct	Numărul de ori de eșec distincte.
distinct	Vector de timpi de eșec distincti.
starea 0	0 dacă estimarea poate fi furnizată; 1 dacă nu, în funcție de dacă alfa ar putea fi estimată.
alpha	Parametrul modelului de fragilitate gamma.
lambda	Estimări ale probabilităților de pericol la momente diferite de defecțiune.
survfunc tvals	Vector de supraviețuire estimat în timpuri distincte.
MLEAttvals	Copie a argumentului.
	Vector de supraviețuire estimat în tvals times.

Următorul segment de cod R poate fi utilizat pentru a calcula un studiu de cohortă pe baza setului de date disponibil:

```
> install.packages(„survrec")
> bibliotecă (survrec)
```

Se încarcă pachetul necesar: boot

Pachetul atașat: „boot"
Următoarele obiecte sunt mascate din „pachet:latice": melanom
Următoarele obiecte sunt mascate din „pachet:supraviețuire": aml

```
> ls("pachet:survrec")
[1] „is.Survr" „mlefrailty.fit" „psh.fit" „q.search"
[5] „surv.search" „survdiffr" „survfitr" „Survr"
[9] "wc.fit"
> date (MMC)
> atașați (MMC)
```

Următoarele obiecte sunt mascate din „colon":
id, timp
Următoarele obiecte sunt mascate din „vezica urinară (poziția 15)":
eveniment, id
Următoarele obiecte sunt mascate din „vezica urinară (poziția 16)":
eveniment, id
Următoarele obiecte sunt mascate din „plămân": timpul

```
> MMC # Afișează datele a 99 de subiecți de caz.
```

	id	timp	Eveniment	grup
1	1	112		Masculii
2	1	145	1	Masculii
3	1	39	1	Masculii
97	19	66	1	Femele
98	19	100	1	Femele
99	19	4	0	Femele

```
>
> fit <- mlefrailty.fit(Survr(MMC$id,MMC$time,MMC$event))
Trebuie să se determine o valoare de semințe pentru Alpha
Seed Alpha: 20,02853
Estimarea alfa= 10,17623
> potrivi
> complot (potrivire)
```

(a) Care sunt funcțiile acestui segment de cod R utilizate în analiza setului de date MMC ?
(b) Rulați segmentul de cod precedent într-un mediu R.
(c) Descrie rezultatele.

REFERINȚE

Doll, R., Peto, R., Boreham, J. și Sutherland, I. (2004). Mortalitatea în legătură cu fumatul: observații de 50 de ani asupra medicilor britanici de sex masculin. *British Medical Journal* , *328* , 1519–1528.

Everitt, BS și Hothorn, T. (2006). *Un manual de analiză statistică folosind R (HSAUR).* Boca Raton, FL: Chapman & Hall/CRC.

Handysides, AR și Landless, PN (2012, februarie). Boli la alegere. *Lumea adventistă — NAD, 8* (2), 19.

Tammemagi, CM, Neslund-Dudas, C., Simoff, M., & Kvale, P. (2004). Fumatul și supraviețuirea cancerului pulmonar: rolul

comorbidității și al tratamentului. *Piept , 125* (1), 27–37.

SEVEN

Studii randomizate, dezvoltare de fază, confuzie în analiza supraviețuirii și regresii logistice

7.1 Încercări aleatorie (STANLEY, 2007)[XXV]

Un **studiu randomizat (RT)** sau **un studiu randomizat controlat (RCT)** este un tip specific de experiment științific și este designul preferat pentru un studiu clinic în - miologia epidemiilor. RT-urile sunt utilizate pentru a testa eficacitatea diferitelor tipuri de intervenții în cadrul unei populații caz-subiect. Ele pot oferi, de asemenea, o oportunitate de a aduna informații utile despre efectele adverse, cum ar fi reacțiile la medicamente.

Trăsătura-cheie distinctivă a RT obișnuită este că subiecții cazului de studiu, *după* evaluarea eligibilității și recrutarea, dar *înainte de* începerea intervenției care urmează să fie studiată, sunt alocați aleatoriu pentru a primi unul sau altul dintre tratamentele alternative din studiu. Alocarea aleatorie în încercările reale este complexă, dar, conceptual , procesul este ca și cum ai arunca o monedă. *După randomizare, cele două grupuri de subiecți sunt urmărite exact în același mod* ; singurele diferențe între îngrijirea pe care o primesc (de exemplu, în ceea ce privește procedurile, testele, vizitele în ambulatoriu, apelurile de urmărire etc.) ar trebui să fie cele intrinsece tratamentelor comparate. Cel mai important avantaj al randomizării adecvate este că minimizează distorsiunea de alocare în alocarea tratamentelor , echilibrând astfel atât factorii de prognostic cunoscuți, cât și necunoscuți.

Clasificări ale RT-urilor după designul studiului

O modalitate de a clasifica RT este prin proiectarea studiului, în care cele patru categorii majore sunt:

Grup paralel — în care fiecare participant este repartizat aleatoriu unui grup și toți participanții din grup primesc (sau nu primesc) o intervenție.

Crossover — în care, în timp, fiecare participant primește (sau nu primește) o intervenție într-o secvență aleatorie.

Cluster — în care grupurile preexistente de participanți (de exemplu, orașe, asociații sociale) sunt selectate aleatoriu pentru a primi (sau nu) o intervenție.

Factorială — în care fiecare participant este repartizat aleatoriu unui grup care primește (sau nu primește) o anumită combinație de intervenții. De exemplu:

- Grupa A primește vitamina X și vitamina Y.

[XXV] Studiu randomizat controlat (RCT). Preluat de la http://en.wikipedia.org/wiki/Randomized _probă_controlată.

- Grupul B primește vitamina X și placebo Y.

- Grupul C primește placebo X și vitamina Y.
- Grupa D primește Placebo X și Placebo Y.

O analiză a celor 616 RT-uri indexate în *PubMed* în decembrie 2006 a constatat că 78% au fost studii în grup paralel, 16% au fost încrucișate, 2% au fost divizate, 2% au fost grupate și 2% au fost factoriale. [1]

RT-urile pot fi, de asemenea, clasificate după *eficacitate* (eficacitatea testului) sau după *ipoteză (* superioritate vs. non-inferioritate vs. echivalență, în funcție de semnificația statistică corespunzătoare).

Randomizare

Următoarele sunt toate avantajele randomizării corecte într-un RT:

- Elimină părtinirea în repartizarea tratamentului , în special părtinirea selecției și confundarea .
- Facilitează orbirea (mascarea) identității tratamentelor de la investigatori , participanți și evaluatori.
- Permite utilizarea teoriei probabilităților pentru a exprima probabilitatea ca orice diferență de rezultat între grupurile de tratament să indice doar șansa.

La randomizarea subiecților de caz, se poate alege dintre două procese:

1. Alegeți o *procedură de randomizare* pentru a genera o secvență imprevizibilă de alocări ; aceasta poate fi o simplă repartizare aleatorie a pacienților la oricare dintre grupuri cu probabilități egale.
2. Alegeți *ascunderea alocării* , care se referă la măsurile de precauție stricte luate pentru a vă asigura că repartizările de grup ale subiecților cazului nu sunt dezvăluite înainte de a fi alocate definitiv grupurilor lor respective.

Metodele „sistematice” non-aleatorie de atribuire a grupului, cum ar fi alternarea subiecților între un grup și altul, pot cauza posibilități nelimitate de contaminare și o încălcare a ascunderii alocării.

PROCEDURI DE RANDOMIZARE

O procedură ideală de randomizare atinge următoarele obiective:

- *Dimensiuni egale ale grupului.* Acest lucru asigură o putere statistică adecvată , în special în analizele de subgrup.
- *Prejudecată scăzută de selecție.* Procedura nu ar trebui să permită unui investigator să prezică alocarea grupului următor subiect prin examinarea cărui grup i s-a repartizat cei mai puțini subiecți până la acel moment .
- *Probabilitate scăzută de confuzie.* Aceasta implică un echilibru al covariatelor între grupuri.

Nicio procedură de randomizare nu poate îndeplini toate aceste obiective în orice împrejurare, așa că, pentru o anumită investigație, epidemiologii ar trebui să aleagă o procedură în funcție de meritele acesteia și de natura investigației.

RANDOMIZARE SIMPLĂ

Randomizarea simplă este o procedură intuitivă folosită în mod obișnuit, similară

aruncării repetate a monedelor corecte; este cunoscută și ca randomizare *completă* sau *nerestricționată* . Este **robust** atât împotriva selecției, cât și a prejudecăților accidentale. Principalul său dezavantaj este posibilitatea unor dimensiuni dezechilibrate ale grupurilor în RT mici. Este recomandat doar pentru RT cu mai mult de 200 de subiecți.

RANDOMIZARE RESTRICȚĂ

se recomandă o formă de *randomizare restricționată* . Unele tipuri majore de randomizare restricționată utilizate în RT sunt:

- **Randomizare cu blocuri permutate** , în care sunt specificate o dimensiune a blocului și un raport de alocare (numărul de subiecți dintr-un grup față de celălalt grup), iar subiecții sunt alocați aleatoriu în cadrul fiecărui bloc. De exemplu, o dimensiune a blocului de 16 și un raport de alocare de 3:1 ar duce la repartizarea aleatorie a 12 subiecți într-un grup și 4 în celălalt. Acest tip de randomizare poate fi combinat cu randomizarea stratificată (de exemplu, pe centru într-un studiu multicentric) pentru a asigura un echilibru bun al caracteristicilor participanților în fiecare grup. Un caz special de randomizare cu blocuri permutate este *alocarea aleatorie* , în care întregul eșantion este tratat ca un singur bloc. Un dezavantaj al randomizării blocurilor permutate este că, chiar dacă dimensiunile blocurilor sunt mari și variate aleatoriu, procedura poate cădea totuși pradă părtinirii selecției. Un alt dezavantaj este că o analiză adecvată a datelor din RCT cu blocuri permutate necesită stratificarea pe blocuri.
- **Randomizare adaptativă cu monede părtinitoare** , în care probabilitatea de a fi repartizată unui grup scade dacă grupul este suprareprezentat și crește dacă grupul este subreprezentat. Se crede că metodele sunt mai puțin afectate de părtinirea selecției decât randomizarea cu blocuri permutate.

ÎNCERCĂRI ALEATORIZATE CU ORBIRE

Un RT poate fi *orbit* (sau *mascat*) prin restricționarea procedurilor utilizate la cele care împiedică subiecții cazului de studiu, îngrijitorii, evaluatorii de rezultate și toți ceilalți care participă sau sunt implicați în studiu să știe care intervenție a fost primită. Cu toate acestea, spre deosebire de ascunderea alocării, orbirea poate fi uneori inadecvată sau imposibil de efectuat într-un RT. De exemplu, dacă un RT implică un tratament în care este necesară participarea activă a subiectului cazului (de exemplu, kinetoterapie), participanții nu pot fi orbiți la intervenție.

Din punct de vedere istoric, RT orb a fost clasificat ca single-orb, dublu-orb sau triplu-orb. În prezent, se preferă ca aceste categorii suplimentare să fie evitate; din 2010, autorii și editorii care raportează RT orb au fost direcționați să discute despre asta după cum urmează: „Dacă se folosește orbirea, atunci raportul ar trebui să identifice cine [a fost] orb după atribuirea intervențiilor (de exemplu, participanții, furnizorii de îngrijire, cei care evaluează rezultatele).) și cum a fost întreprinsă orbirea” (Blanchard, 2012; Moreira , Araujo și Machado, 2012; Sigler & Stemhagen, 2011).

RCT-urile fără orbire sunt denumite *neorb* sau deschise. În 2008, un studiu a concluzionat că rezultatele RT neorb tind să fie părtinitoare către efecte benefice numai dacă rezultatele erau mai degrabă subiective decât obiective. De exemplu, într-un RT de tratamente pentru scleroza multiplă, neurologii neorbiți (dar nu neurologii orbiți) au considerat că tratamentele au fost benefice. În practică, deși participanții și furnizorii implicați într-un RT sunt adesea neobișnuiți, este de dorit și adesea posibil să-l orbiți pe

evaluator sau să obțineți o sursă obiectivă pentru a evalua rezultatele.

Analiza biostatistică a datelor din RT

Tipurile de metode statistice utilizate în RT depind de caracteristicile datelor.

- Pentru **datele de rezultat dihotomice (binare)** , se pot utiliza regresia logistică și alte metode (de exemplu, pentru a prezice răspunsuri virusologice susținute după primirea tratamentului pentru hepatita C).
- Pentru **date continue de rezultat** , analiza covarianței poate fi utilizată pentru a testa efectele variabilelor predictoare (de exemplu, pentru modificări ale nivelurilor lipidelor din sânge după primirea tratamentului pentru sindromul coronarian acut) .
- Pentru **datele de rezultat din timpul până la eveniment** care pot fi cenzurate, analiza de supraviețuire este adecvată (de exemplu, estimatorii Kaplan–Meier [K–M] și modelele de riscuri proporționale Cox pentru timpul până la boala coronariană [CHD] după primirea terapiei de substituție hormonală în menopauză).

Biostatistică pentru RT în mediul R

Pe site-ul web CRAN, suportul pentru studiile clinice randomizate este disponibil în ceea ce privește proiectarea, monitorizarea și analiza proiectului. Pentru a accesa aceste surse din Mediu :

Selectați pagina de start Ajutor/CRAN
Selectați opțiunea „Căutare/Căutare în site-ul R ” și introduceți „Studii clinice randomizate”

Această selecție apelează multe surse disponibile, inclusiv următoarele.

- 1) Vizualizarea activității CRAN: Proiectare, monitorizare și analiză a studiilor clinice (scor: 3) ***Autor:*** Necunoscut
  Data: miercuri, 25 apr 2012 02:57:32
  CRAN Task View: Pachete CRAN de proiectare, monitorizare și analiză a studiilor clinice
  Legături înrudite: ClinicalTrials menținerea informațiilor de vizualizare a activității: Ed Zhang *Contact:* Ed.Zhang.jr@gmail.com
  Versiunea: 201

De la http://finzi.psych.upenn.edu/views/ClinicalTrials.html (23.531 de octeți), se pot selecta suporturile metodologice adecvate pentru biostatistica RT, inclusiv proiectarea, monitorizarea și analiza proiectului. Pachetele CRAN sunt disponibile în următoarele domenii ale RT:

- Proiectare si monitorizare
- Proiectare și analiză
- Analiza pentru proiecte specifice
- Analiza in general
- Meta-analiză

Următoarele exemple sunt tipice.

■ **Exemplul 7.1:** Proiectarea RT-urilor utilizând dimensiunea eșantionului de pachet : Determinarea dimensiunii eșantionului

Pachetul CRAN dimensiunea eșantionului calculează dimensiunea eșantionului pentru testul *t* Student , cu variații egale și neegale; și pentru testul Wilcoxon–Mann–Whitney pentru date categorice, cu și fără legături.

t Student independent cu dimensiunea grupului inegală, utilizați funcția n.indep.t.test.eq() , cu parametrii impliciti:

n.indep.t.test.neq(putere = 0,8, alfa = 0,95, medie.dif = 0,8, sd.est = 0,83, k = 0,5)

controlat de următorii parametri variabili ca argumente:

puterea necesara putere 1-beta
alpha necesar Nivel I-eroare 1-alpha
medie.dif diferența minimă necesară între mediile grupului
sd.est abaterea standard pe grupuri
k n_2 = n_1*k
n_1 dimensiunea eșantionului din grupul 1
n_2 dimensiunea eșantionului din grupul 2
N dimensiune totală a eșantionului: N=n_1 + n_2

Următorul segment de cod R poate fi utilizat pentru a calcula dimensiunile eșantioanelor pentru un studiu clinic randomizat folosind acest pachet CRAN:

■ install.packages(„mărimea eșantionului")
■ bibliotecă (dimensiune eșantion)
■ ls(„pachet:dimensiune eșantion")

[1] „n.indep.t.test.eq" „n.indep.t.test.neq" „n.paired.t.test"
[4] „n.welch.test" „n.wilcox.ord"

```
> n.indep.t.test.eq(putere = 0,8, alfa = 0,95, medie.dif = 0,8,
+                   sd.est = 0,83) # Pentru două mostre de dimensiuni egale:
[1] "sample.size:" "29"
>  # Mergând pentru un test cu o putere biostatistică mare, să zicem , putere = 0,95
>  n.indep.t.test.eq(putere = 0,95, alfa = 0,95, medie.dif = 0,8,
+                   sd.est = 0,83)
[1] "sample.size:" "49"
>  # Și pentru două mostre de dimensiuni inegale
>  n.indep.t.test.neq(putere = 0,8, alfa = 0,95, medie.dif = 0,8,
+ sd.est = 0,83) # La 0,8 putere
[1] „dimensiune.probă:" „32" „dimensiune.probă nr.1:" „21,3"
[5] „dimensiune.probă nr.2:" „10,7"

> n.indep.t.test.neq(putere = 0,95, alfa = 0,95, medie.dif = 0,8,
+ sd.est = 0,83) # Și la 0,95 putere
[1] „dimensiune.probă:" „55" „dimensiune.probă nr.1:" „36,7"
[5] „dimensiune.probă nr.2:" „18,3"
```

Remarcă:
Pentru o putere biostatistică mai mare, sunt necesare probe mai mari.

Întrebări de revizuire pentru Secțiunea 7.1

1. (a) Ce este un RT în contextul unei investigații epidemiologice?
   (b) Care sunt avantajele RT într-un studiu precum un studiu clinic?
2. (a) Ce este randomizarea restricționată?
   (b) Explicați următorii termeni: randomizare cu bloc permutat, randomizare adaptativă cu monede părtinitoare.
3. În analiza biostatistică a datelor din RT, sunt de așteptat trei caracteristici ale datelor. Numiți-le și explicați pe scurt fiecare caracteristică.
4. Pachetele CRAN sunt disponibile pentru a analiza datele de la RT. În ce cinci domenii ale RT sunt aplicabile aceste pachete?

Exerciții pentru Secțiunea 7.1

1. Pentru exemplul lucrat 7.1,
   (a) Explicați funcția fiecărei linii a segmentului de cod R pentru calcul.
   (b) Reluați acest segment de cod în mediul R.
2. În pachetul CRAN CRTSize , funcția n4means() poate fi utilizată pentru a furniza informații despre estimarea dimensiunii eșantionului. De exemplu, poate calcula numărul de subiecte de caz necesare pentru un cluster RT cu rezultat continuu. Următorul segment de cod R este utilizat atunci când rezultatul este continuu (de exemplu, tensiunea arterială sau greutatea). Rețineți că, dacă rezultatele sugerează că este necesar un număr mic de clustere, o procedură iterativă va include distribuția *t* în loc de valoarea critică normală pentru alfa, iterând până la convergență. Pentru această funcție n4means() , se aplică următoarea specificație:

Descriere:
Această funcție oferă informații detaliate de estimare a dimensiunii eșantionului pentru a determina numărul de subiecți care trebuie înscriși într-un grup RT pentru a compara două medii.

Utilizare:
n4 înseamnă (delta, sigma, m, ICC, alfa=0,05, putere=0,8, AR=1, two.tailed=TRUE, cifre=3)

Argumente:

delta	Diferența minimă detectabilă între mediile populației. sigma Eroarea standard a rezultatului.
m	Dimensiunea medie anticipată (sau reală) a clusterului.
ICC	Valoarea anticipată a coeficientului de corelație intraclasă, *r* .
RA	Raportul de alocare: AR=1 implică un număr egal de subiecți per grup de tratament și control (eficiență maximă), AR > 1 implică

că mai mulți subiecți vor fi înscriși în grupul de control (de exemplu, în cazul unei intervenții costisitoare), iar AR < 1 implică că vor fi înscriși mai mulți subiecți în grupul de tratament (folosit rar).

alpha Rata de eroare de tip I dorită.

putere Nivelul dorit de putere, recall putere = 1 - Eroare de tip II.

două.cozi Valoare logică. Dacă TRUE , calculele se bazează pe o eroare de tip I cu două cozi; dacă FALSE , se efectuează un calcul unilateral.

cifre Numărul de cifre pentru a rotunji calculele.

Valoare:

nE Numărul minim de subiecți necesar în lotul experimental.

nC Numărul minim de subiecți necesar în grupul de control.

delta Diferența minimă detectabilă între mediile populației. sigma Eroarea standard a rezultatului.

alpha Rata de eroare de tip I dorită.

putere Nivelul dorit de putere, recall putere = 1 - Eroare de tip II.

AR Raportul de alocare.

Următorul segment de cod R este disponibil pentru a efectua calculul:

```
> install.packages("CRTSize")
> library(CRTSize)
> ls("package:CRTSize")
[1] "fixedMetaAnalMD"                "fixedMetaAnalRROR"
[3] "n4incidence"                    "n4means"
[5] "n4meansEB"                      "n4meansMeta"
[7] "n4props"                        "n4propsEB"
[9] "n4propsMeta"                    "print.fixedMetaAnalMD"
[11]"print.fixedMetaAnalRROR"        "print.n4incidence"
[13] "print.n4means"                 "print.n4meansEB"
[15] "print.n4meansMeta"             "print.n4props"
[17] "print.n4propsEB"               "print.n4propsMeta"
[19] "summary.fixedMetaAnalMD"       "summary.fixedMetaAnalRROR
[21] "summary.n4incidence"           "summary.n4means"
[23] "summary.n4meansEB"             "summary.n4meansMeta"
[25] "summary.n4props"               "summary.n4propsEB"
[27] "summary.n4propsMeta"
```

```
> n4 înseamnă (delta=10, sigma=1, m=25, ICC=0,05,
+             alfa=0,05, putere=0,80);
> # Ieşire:
```

Dimensiunea necesară a eșantionului este de minim 1 grup de mărime 25 în Grupul Experimental și de minim 1 grup (dimensiune 25) în Grupul de control.

(a) Explicați funcția fiecărei linii a segmentului de cod R pentru acest calcul.

(b) Reluați acest segment de cod în mediul R.

(c) Recalculați estimarea dimensiunilor clusterului pentru o putere biostatistică de 0,90.
(d) Comentați rezultatele.

3. Pachetul CRAN randomSurvivalForest descrie pădurile de supraviețuire aleatoare pentru date de supraviețuire cu cenzură la dreapta și riscuri concurente (Gerds, Cai și Schum acher, 2008; Graf, Schmoor, Sauerbrei și Schumacher, 1999) .

Rezultatele funcției plot.ensemble() din acest pachet sunt curbe de supraviețuire ansamblu și estimări de ansamblu ale mortalității.

Această abordare este aplicabilă analizelor de risc concurente, dar parcelele nu sunt specifice unui eveniment. Pentru curbele specifice evenimentului și pentru o analiză mai cuprinzătoare, utilizați competing.risk în astfel de cazuri.

Următorul segment de cod R este disponibil pentru a efectua calculul, folosind veteranul setului de date în supraviețuirea pachetului CRAN :

```
> install.packages(„randomSurvivalForest")
> bibliotecă (randomSurvivalForest)
randomSurvivalForest 3.6.3
Tastați rsf.news() pentru a vedea funcții noi, modificări și remedieri de erori.
> ls(pachet:randomSurvivalForest) # Listarea conținutului:

[1] „risc.concurent"          „găsește.interacțiu     "imputare.rsf"
[4] „max.subtree"             "complot.ansambl        "plot.error"
[7] „plot.proximity"          "plot.rsf"              "plot.variable"
[10] „pmml2rsf"               „predict.rsf"           „print.rsf"
[13] „randomSurvivalForest"   "rsf"                   "rsf.news"
[16] „rsf2pmml"               "rsf2rfz"               "varSel"
[19] „vimp"

Mesaj de avertizare:
În ls(pachet:randomSurvivalForest):
„pachet:randomSurvivalForest" convertit în șir de caractere
> install.packages(„supraviețuire")
> bibliotecă (supraviețuire)
Se încarcă pachetul necesar: spline
> ls("pachet:supraviețuire") # Listarea conținutului:

[1] „aareg"                   "aml"                   "attrassign"
[4] „basehaz"                 "vezica urinara"        "vezica 1"
[7] „vezica urinară2"         "cancer"                "cch"
[10] „cgd"                    "clogit"                "cluster"
[13] „colon"                  „cox.zph"               "coxph"
[16] „coxph.control"          "coxph.detail"          "coxph.fit"
[19] „dsurvreg"               "format.Surv"           "fragilitate"
[22] „frailty.gamma"          "fragilitate.gaussian"  "fragilitate.t"
[25] „inima"                  „este.na.coxph.penalt   „este.na.evalu
```

[28] „is.na.Surv" **"este.evaluabil"** **"este.Surv"**

[31] „jasa" **"jasa1"** **"rinichi"**
[34] „labels.survreg" **"leucemie"** **"Logan"**

[37] „plămân" **"match.ratetable"** **"mgus"**

[40] „mgus1" **"mgus2"** **"nwtco"**

[43] „ovarian" **"pbc"** **"pbcseq"**
[46] „pspline" **"psurvreg"** **"pears"**

[49] „qsurvreg" **"tabel de evaluare"** **"evaluareData"**

[52] „șobolani" **"cresta"** **"stanford2"**
[55] „strat" **„Surv"** **„survConcordanc**

[58] „survdiff" **"survexp"** **"survexp.mn"**
[61] „survexp.us" **"survexp.usr"** **"survfit"**

[64] „survfitcoxph.fit" **"survobrien"** **"survreg"**

[67] „survreg.control" **"survreg.distributions"** **"survreg.fit"**
[70] „survregDtest" **"survSplit"** **"tcut"**

[73] „tobin" **"tt"** **"descurca.special**

[76] „veteran"

```
> date(veteran, pachet = "randomSurvivalForest")
> v.out <- rsf(Surv(time, status) ~ ., veteran, ntree = 1000)
> date(veteran, pachet = "randomSurvivalForest")
> v.out <- rsf(Surv(time, status) ~ ., veteran, ntree = 1000)
> complot.ansamblu (v.out)
> # Ieşire: Figura 7.1 : randomSurvivalForest -1
```

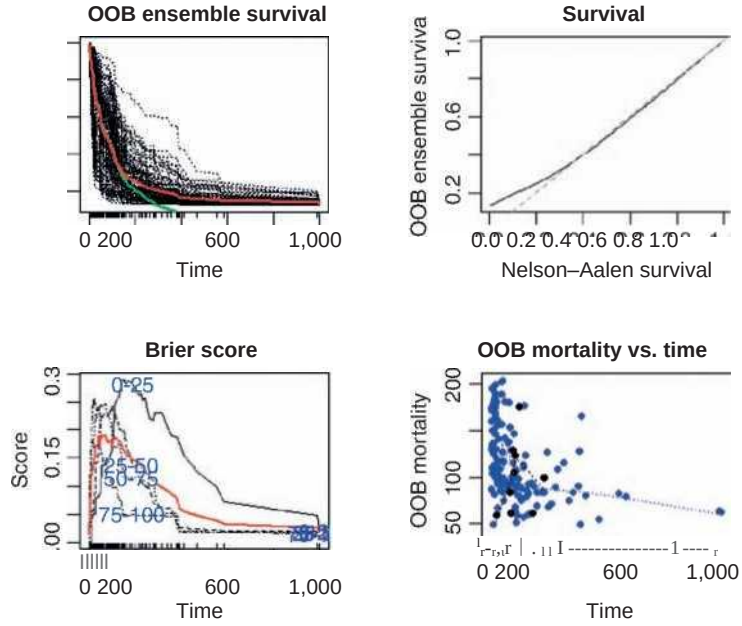


FIGURE 7.1 randomSurvivalForest-1.

Sunt produse patru parcele (*de sus în jos, de la stânga la dreapta*):

(1) Funcția de supraviețuire a ansamblului pentru fiecare individ. Linia roșie groasă este supraviețuirea generală a ansamblului; linia groasă verde este estimatorul Nelson–Aalen.
(2) Comparația funcției de supraviețuire a ansamblului populației cu estimatorul Nelson–Aalen.
(3) Scorul Brier (0 = perfect, 1 = slab și 0,25 = ghicit) stratificat în funcție de mortalitatea ansamblului. Pe baza metodei descrise în Gerds et al. (2008) , în care distribuția de cenzură este estimată folosind estimatorul K–M. Stratificarea este în patru grupe corespunzătoare valorilor de mortalitate 0–25, 25–50, 50–75 și 75–100 procente ale mortalității. Linia roșie este scorul general (nestratificat) Brier.
(4) Graficul mortalității în funcție de timpul observat. Punctele în albastru corespund evenimentelor; punctele cu negru sunt observații cenzurate.

```
> plot(v.out) # Ieşire: Figura 7.2 : randomSurvivalForest -2
```

	Importanță	Relativ Imp
karno	**0,1410**	**1,0000**
tipul	**0,0405**	**0,2871**
trt	**0,0036**	**0,0257**
vârstă	**0,0001**	**0,0008**
anterior	**-0,0018**	**-0,0128**
diagtime	**-0,0019**	**-0,0136**

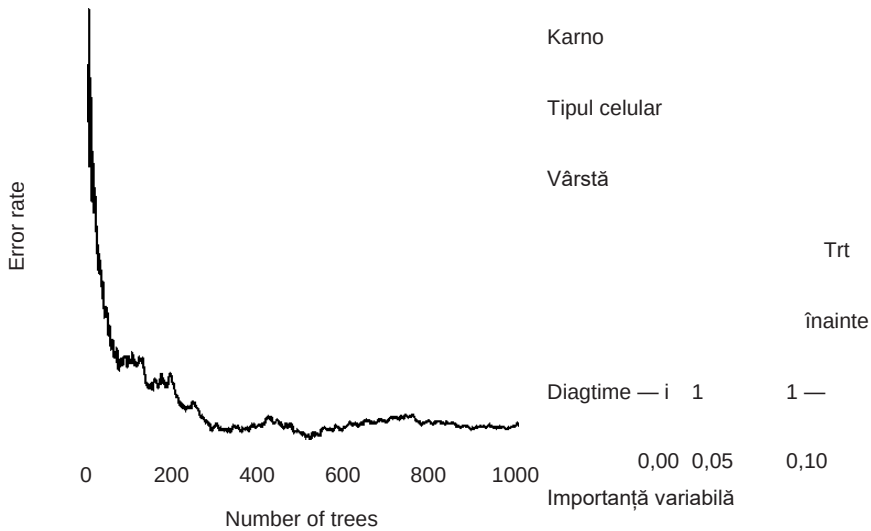


FIGURA 7.2 randomSurvivalForest -2.

```
> # complot de supraviețuire a ansamblului pentru un singur individ
> surv.ensb <- t(exp(-v.out$oob.ensemble))
> plot(v.out$timeInterest, surv.ensb[, 1])
> # Ieşire: Figura 7.3 : randomSurvivalForest -3
>
```

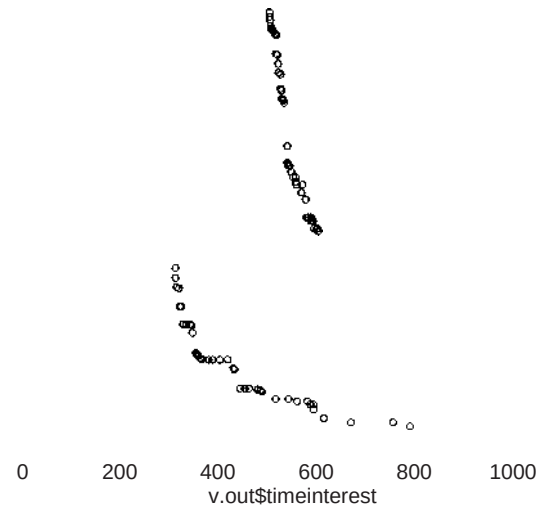


FIGURA 7.3 randomSurvivalForest -3: graficul supraviețuirii ansamblului pentru un singur individ.

(a) Explicați funcția fiecărei linii a segmentului de cod R pentru acest calcul.
(b) Reluați acest segment de cod în mediul R.
(c) Recalculați estimarea dimensiunilor clusterului pentru un alt set de date din supraviețuirea pachetului CRAN . Comentează rezultatele.

4. Pentru a obține efectele de supraviețuire ale variabilelor, funcția R plot.variable() , din pachetul CRAN randomSurvivalForest (rSF), poate fi utilizată pentru a crea grafice ale mortalității ansamblului, supraviețuirii estimate sau timpului de supraviețuire prezis față de o variabilă *x dată.* . Pot fi create și parcele marginale și parțiale. Fie mortalitatea , frecvența relativă a mortalității, supraviețuirea prognozată, fie timpii de supraviețuire estimați sunt reprezentați grafic pe axa verticală (valoarea *y*) în funcție de variabilele *x de pe* axa orizontală.
   - Alegerea lui *x* variabilele pot fi specificate folosind nume de predictori.
   - Alegerea valorii *y* este controlată de tip.
     Există patru opțiuni diferite:
     (1) „mort" este mortalitatea de ansamblu.
     (2) „rel.freq" este mortalitatea standardizată.
     (3) „surv" este supraviețuirea prezisă la un moment dat.
     (4) „timpul" este timpul estimat de supraviețuire.

Pentru variabilele continue, punctele sunt colorate cu albastru, corespunzând evenimentelor, și negru, pentru observațiile cenzurate. Mortalitatea ansamblului ar trebui interpretată în funcție de numărul total de decese. De exemplu, dacă i are o valoare a mortalității de 100, atunci dacă toți indivizii ar fi la fel ca i , numărul așteptat de decese ar fi 100. Dacă type="rel.freq" , atunci valorile mortalității sunt împărțite la o dimensiune ajustată a eșantionului , definit ca valoarea maximă a dimensiunii eșantionului și valoarea maximă a mortalității. Valorile standardizate ale mortalității nu indică decesele totale, ci mai degrabă o mortalitate relativă.

Graficele parțiale sunt create când partial=TRUE . Interpretările pentru acestea sunt diferite de cele pentru parcelele marginale. Valoarea parțială pentru o variabilă X, evaluată la X = x , este

$$f(x) = (1/n)\sum f(x; X_{i,\varnothing})_{i=1}$$

unde *f* este valoarea prezisă; și, pentru fiecare individ i , $x_{i,o}$ reprezintă valoarea pentru toate celelalte variabile, altele decât X . Pentru variabilele continue, punctele roșii sunt folosite pentru a indica valorile parțiale ; liniile roșii întrerupte reprezintă o bară de eroare de +/- două erori standard. O linie neagră întreruptă indică cea mai mică estimare a valorilor parțiale. Pentru variabilele discrete , valorile parțiale sunt indicate folosind diagrame de casete cu mustăți care extind aproximativ două erori standard de la medie. Erorile standard sunt furnizate doar ca ghid și trebuie interpretate cu prudență. Parcele parțiale pot fi lente. Setarea type="time" poate îmbunătăți lucrurile. De asemenea, ar trebui să încercați să setați npts la un număr mai mic.

Pentru analizele de risc concurente, graficele corespund unor valori necondiționate (adică, nu sunt specifice unui eveniment). Utilizați competiting.risk pentru curbele specifice evenimentului și pentru o analiză mai cuprinzătoare în astfel de cazuri.

Formula de utilizare pentru această funcție este

```
plot.variable(x, plots.per.page = 4, granule = 5, sorted = TRUE, type = c("mort",
              "rel.freq", "surv", "time")[1], parțial = FALSE, predictorNames = NULL,
              npred = NULL, npts = 25, subset = NULL, percentila = 50, …)
```

în care argumentele sunt

x Un obiect de clasă (rsf, grow) sau (rsf, predict).
plots.per.page Valoare întreagă care controlează aspectul paginii.
granule Valoare întreagă care controlează dacă o diagramă pentru o anumită variabilă ar trebui să fie dată ca diagramă cu box sau diagramă de dispersie. Valorile mai mari forțează boxplots.
sortate Variabilele trebuie sortate după valori de importanță (se aplică numai dacă sunt disponibile valori de importanță)?
tip Selectați tipul de valoare care trebuie trasată pe axa verticală.
parțial Ar trebui să fie create parcele parțiale?
predictorNames Vector de caractere al *x* variabilelor de reprezentat grafic. Implicit este tot.
npred Numărul de variabile de reprezentat grafic. Implicit este tot.
npts Numărul maxim de puncte utilizate la generarea graficelor parțiale pentru variabile continue .
subset Indici care indică ce rânduri ale matricei de predictori vor fi utilizate. (Notă: Acest lucru se aplică matricei de predictori procesate, predictorilor obiectului.) Implicit este utilizarea tuturor rândurilor.
percentilă Percentila de timp de urmărire utilizată pentru reprezentarea grafică a supraviețuirii prezise.
... Argumente suplimentare transmise către sau de la alte metode.

Următorul segment de cod R este disponibil pentru a efectua calculul utilizând veteranul setului de date în supraviețuirea pachetului CRAN :

```
> install.packages(„randomSurvivalForest")
> bibliotecă (randomSurvivalForest)

>   ls(pachet: randomSurvivalForest)
>
>   install.packages(„supraviețuire")
>   bibliotecă (supraviețuire)
```

```
>  ls(„pachet:supraviețuire")
>
>  # Câteva exemple aplicate datelor veteranilor .
>  date(veteran, pachet = "randomSurvivalForest")
>  v.out <- rsf(Surv(time,status) ~ ., veteran, nsplit = 10,
+            ntree = 1000)
>
>  plot.variable(v.out, plots.per.page = 3)
```

Mesaj de avertizare:
În bxp(list(stats = c(28.3583161744737, 56.5401467687752, 78.2666123470398): unele crestături au ieșit în afara balamalelor („box"): poate setați notch=FALSE))

```
>  # Ieșire: Figura 7.4 : rSF-1
>
>  plot.variable(v.out, plots.per.page = 2,
+                predictorNames = c("trt", "karno", "varsta"))
>  # Ieșire: Figura 7.5 : rSF-2
>  plot.variable(v.out, type = "surv", npred = 1, percentila = 50)
>  # Ieșire: Figura 7.6 : rSF-3
>  plot.variable(v.out, type = "rel.freq", parțial = TRUE,
+ plots.per.page = 2, npred=3)
>  # Ieșire: Figura 7.7: rSF-4
```

(a) Explicați funcția fiecărei linii a segmentului de cod R pentru acest calcul.
(b) Reluați acest segment de cod în mediul R.
(c) Recalculați estimarea dimensiunilor clusterului pentru un alt set de date din supraviețuirea pachetului CRAN . Comentează rezultatele.

5. Repetați calculul ca în exercițiul 4, dar de data aceasta, utilizați setul de date pbc , după cum urmează:

```
>
>  install.packages(„randomSurvivalForest")
>  bibliotecă (randomSurvivalForest)
>  ls(pachet: randomSurvivalForest)
>
>  install.packages(„supraviețuire")
>  bibliotecă (supraviețuire)
>  ls(„pachet:supraviețuire")
>
>  # Grafice parțiale rapide folosind tipul „timp".
>  # Top 8 predictori din datele PBC
>  date(pbc, package = "randomSurvivalForest")
>  pbc.out <- rsf(Surv(zile,stare) ~ ., pbc, ntree=1000,nsplit=3)
>  plot.variable(pbc.out, type = "time", parțial = TRUE, npred=8)
>  # Ieșire: Figura 7.8 : rSF-5
```

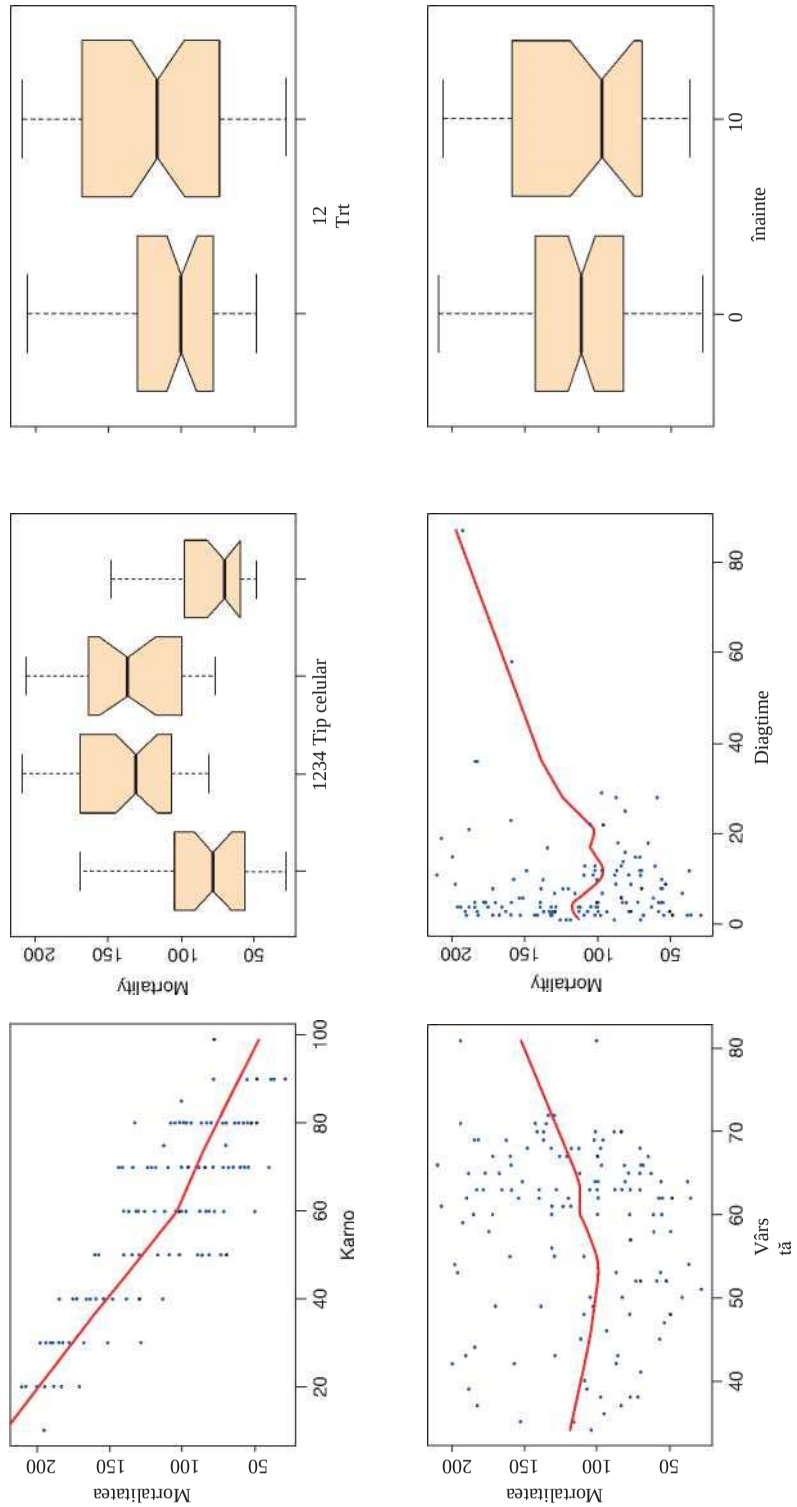


FIGURA 7.4 rSF-1.

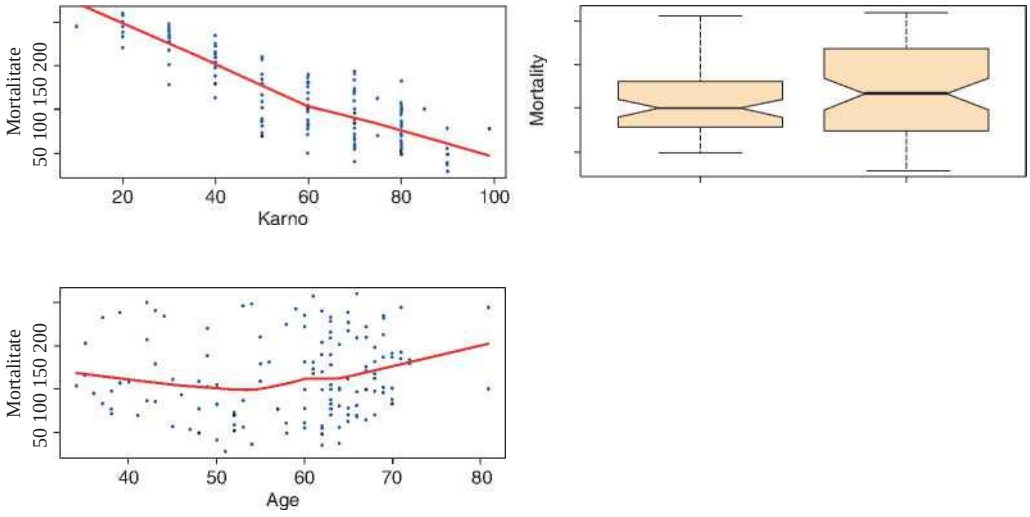


FIGURA 7.5 rSF-2.

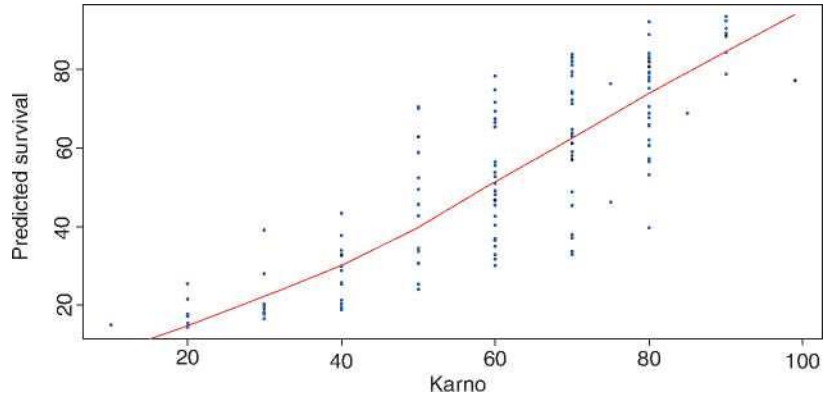


FIGURA 7.6 rSF-3.

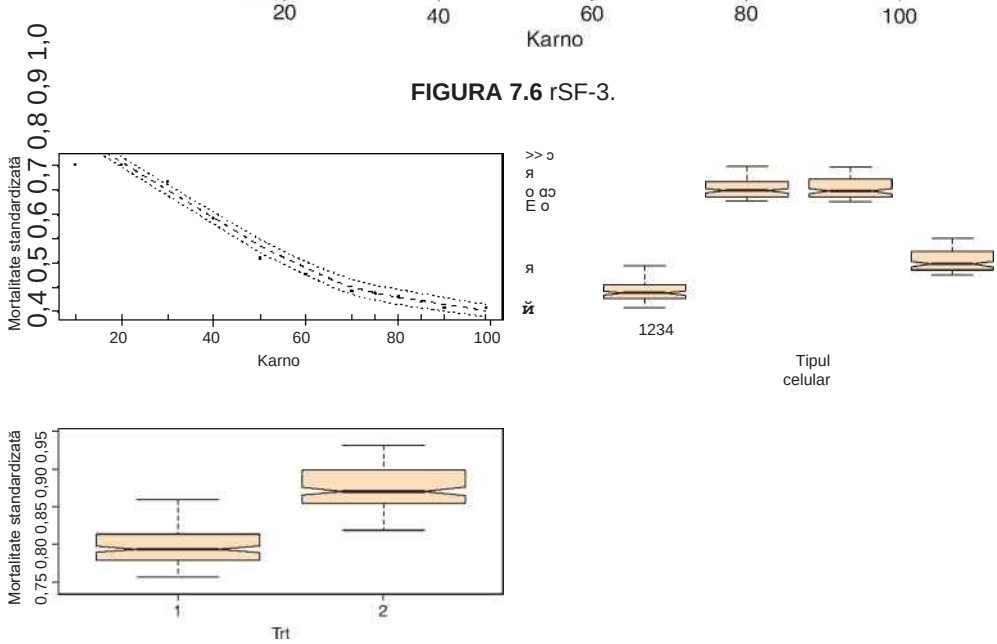


FIGURA 7.7 rSF-4.

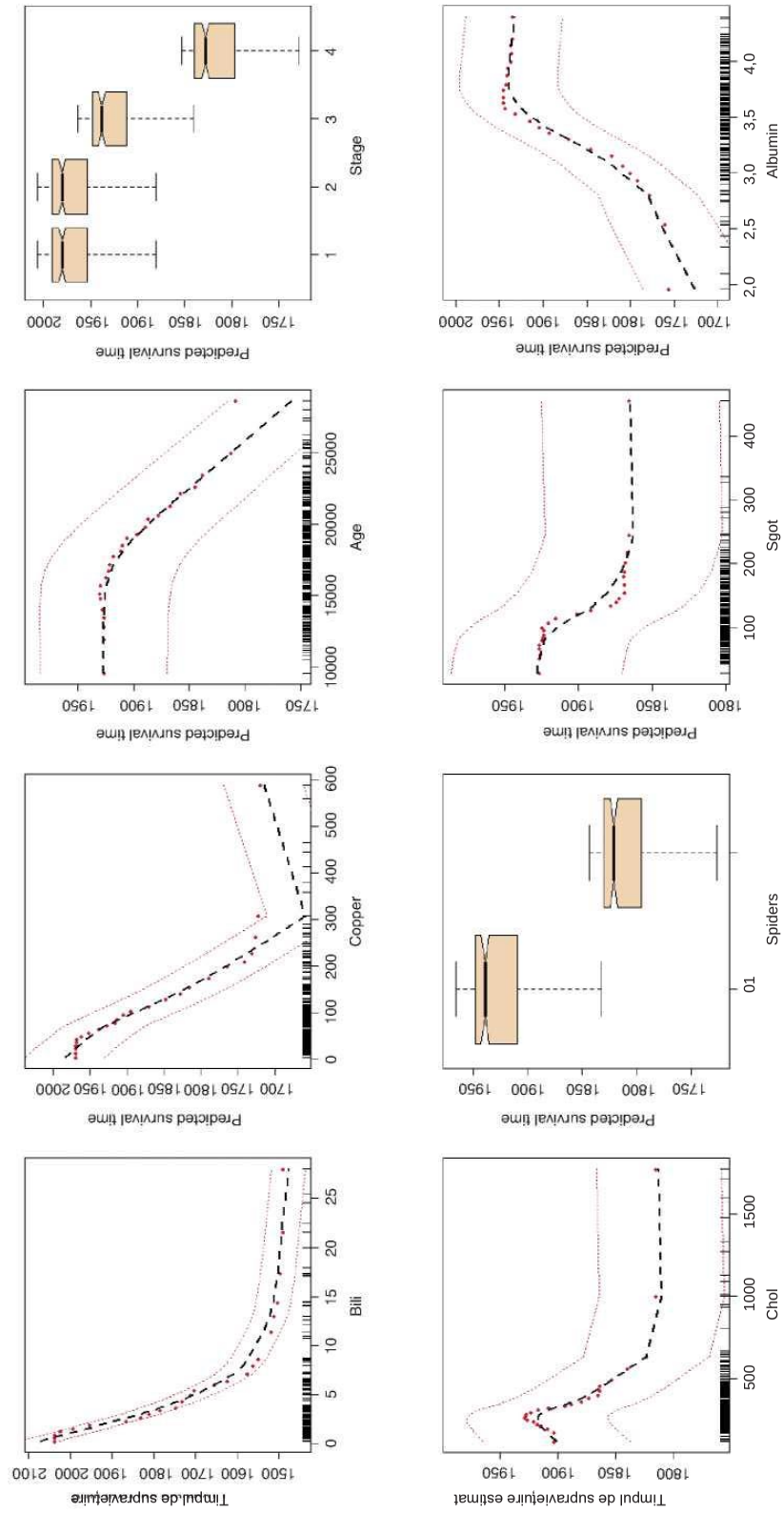


FIGURA 7.8 rSF-5.

(a) Explicați funcția fiecărei linii a segmentului de cod R pentru acest calcul.
(b) Reluați acest segment de cod în mediul R.
(c) Rulați din nou acest segment de cod în mediul R , dar utilizați următoarea comandă de ieșire:

```
> plot.variable(pbc.out, type = "time", parțial = TRUE, npred=16)
```

la ieșire Figura 7.9 : rSF-6. Comentează rezultatele.

7.2 FAZA DE DEZVOLTARE (CALVERT & PLUMMER, 2008; SIGLER & STEMHAGEN, 2011)

În Statele Unite, precum și în multe alte țări din întreaga lume, procesul de dezvoltare a unui nou medicament pentru a satisface o anumită nevoie de sănătate constă din mai multe etape sau faze definitive. De obicei, aceste faze de dezvoltare sunt rezumate în Tabelul 7.1 .

În toate fazele, dimensiunea necorespunzătoare sau inadecvată a eșantionului de studiu este un defect comun de proiectare a studiului clinic.

Faza 0 sau faza preclinica

În această fază inițială, medicamentul este în general testat in vitro (celule, eprubete) sau in vivo (animale). Dezvoltatorul solicită permisiunea guvernamentală pentru a intra în teste clinice. Procedura de solicitare a permisului va depinde de țară. De exemplu, în Statele Unite, trebuie să se acorde o cerere de investigație pentru un nou medicament (IND) înainte de a începe studiile clinice.

Termenul **preclinic** cuprinde toate studiile întreprinse înainte de începerea studiilor clinice. Aceasta include **programe de cercetare** , din care cele mai promițătoare compuși sunt selectați pentru dezvoltare ulterioară.

Conceptul unui studiu de fază 0 este relativ nou. Există acum o tendință în creștere lent de a baza proiectele de studii clinice timpurii pe obiectivele farmacocinetice și farmacodinamice care au fost dezvoltate în investigațiile preclinice. În general, proiectele de studii utilizează obiective farmacodinamice și agenți țintiți.

TABELUL 7.1 Fazele dezvoltării medicamentelor

FAZA SAU ETAPA	OBIECTIVE
Faza 0 sau preclinic	Studii de toxicologie pe animale
Faza I	Studii de farmacologie clinică
Faza II	Studii doză-răspuns
Faza III	Determinarea eficacității și siguranței
Faza IV	Determinarea siguranței produsului pentru utilizare într-o populație

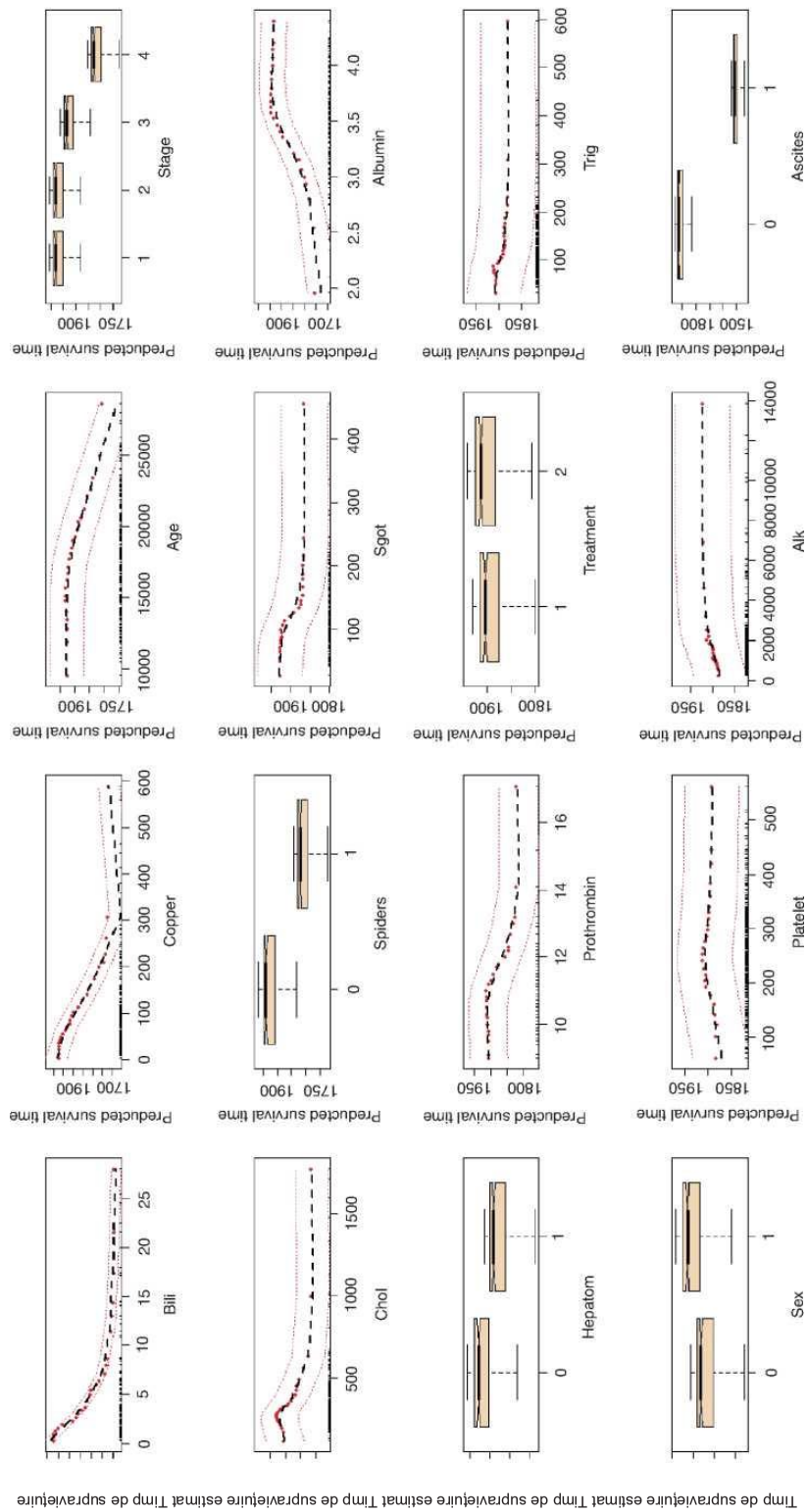


FIGURA 7.9 rSF-6.

Faza I

Obiectivele fazei I includ determinări de bază ale siguranței și eficacității și descoperirea inițială și descrierea farmacocineticii. În această fază, sunt întreprinse primele încercări ale unei noi terapii sau medicamente, de obicei efectuate folosind voluntari bărbați sănătoși. Subiecții de caz (pacienții) pot fi evaluați în locul voluntarilor în studiile de fază I pentru a trata imediat afecțiunile grave și care pun viața în pericol pentru care nu este disponibilă nicio terapie alternativă comparabilă sau satisfăcătoare (în Statele Unite, aceasta se numește *statutul de tratament IND*) . De asemenea, programele de acces extins permit utilizarea subiecților de caz pentru care terapia standard este ineficientă sau contraindicată; în cazul în care astfel de persoane nu sunt eligibile pentru a intra în studii, acestea pot primi medicamente de investigație în paralel cu studiile controlate.

Faza II

Această fază urmărește să ofere o măsură a eficacității în plus față de tolerabilitatea pe termen scurt . Studiile de fază II sunt efectuate în cazul subiecților care au boala sau condiția pe care medicamentul este destinat să o trateze. Mai mult, obiectivele studiului de fază II includ determinarea dozei minime care este eficientă maximă sau care este suficient de eficientă fără toxicitate excesivă.

În unele studii, folosind cunoștințele de cercetare și dezvoltare, un program de fază II poate combina unele studii de fază I și faza II:

> Faza IIA — Teste -pilot sau de fezabilitate
> Faza IIB - Studii pivot, bine controlate

Faza III

Scopul în faza III este de a confirma eficacitatea și de a monitoriza reacțiile adverse din utilizarea pe termen lung. În studiile de fază III, un medicament este testat în condiții care se aseamănă mai mult cu cele în care medicamentul ar fi utilizat dacă ar fi aprobat pentru comercializare. Scopul este de a aduna informații suplimentare despre eficacitate și tolerabilitate care sunt necesare pentru evaluarea relației generale risc-beneficiu a medicamentului și de a oferi o bază adecvată pentru etichetarea medicului.

Deciziile de aprobare/dezaprobare se bazează pe rezultatele unor studii (pivotale) adecvate și bine controlate. Pentru a fi considerat *pivot* , un studiu trebuie să îndeplinească cel puțin următoarele patru criterii definite de Food and Drug Administration (FDA):

> Fii controlat folosind placebo sau o terapie standard .
> design dublu-orb atunci când un astfel de design este practic și etic.
> Fii randomizat.
> Să fie de dimensiuni adecvate .

Farmacoepidemiologie: o ramură a epidemiologiei

Farmacoepidemiologia este studiul utilizărilor și efectelor, atât adverse, cât și benefice , ale medicamentelor în populație. Un exemplu tipic al muncii de farmacoepidemiologie

este Registrul de sarcini cu medicamente antiepileptice (AED), un studiu de cohortă observațional sau un registru al femeilor care rămân însărcinate în timpul tratamentului cu AED. Registrul înregistrează rezultatele nașterii și investighează dacă rezultatele adverse ale nașterii variază în funcție de utilizarea anumitor AED. Deoarece informațiile sunt colectate în primul rând prospectiv, acest studiu este un studiu de cohortă prospectiv.

Un **studiu de cohortă prospectiv** urmărește, de-a lungul timpului, un grup de subiecți de caz similar (*cohorte*) care diferă în ceea ce privește anumiți factori aflați în studiu, pentru a determina modul în care acești factori afectează ratele unui anumit rezultat. De exemplu, s-ar putea urmări o cohortă de șoferi de camion tineri (sub 30 de ani) ale căror obiceiuri de fumat variază, pentru a testa ipoteza că rata de incidență a cancerului pulmonar pe 20 de ani va fi cea mai mare în rândul fumătorilor înrăiți, urmați de fumătorii moderati. , iar apoi nefumătorii . Studiile prospective sunt importante pentru cercetările privind etiologia bolilor și tulburărilor la om deoarece, din motive etice, oamenii nu pot fi expuși în mod deliberat factorilor de risc suspectați în experimente controlate. Studiile de cohortă prospective sunt de obicei clasate mai sus în ierarhia dovezilor decât studiile de cohortă retrospective.

Un **studiu de cohortă retrospectiv** analizează în general evenimentele care au avut loc deja. În medicină, *studiul retrospectiv* se referă de obicei la revizuirea (privirea înapoi la) istoricul medical sau stilul de viață al unui pacient.

Unul dintre avantajele studiilor prospective de cohortă este că pot ajuta la determinarea factorilor de risc pentru a fi infectat cu o nouă boală. Acest lucru este posibil deoarece ele constau în observații longitudinale în timp, iar colectarea rezultatelor se face la intervale regulate de timp, minimizând eroarea de reamintire.[26]

Farmacoepidemiologii trebuie să facă comparații adecvate între pacienții tratați și cei netratați. Printre provocări se numără **distorsiunea de canalizare** , unde grupurile tratate cu Drug x versus Drug y sunt diferite în moduri care prezic rezultatul. Această eroare este o considerație constantă care a determinat dezvoltarea unor abordări cum ar fi scorurile de înclinație pentru a minimiza distorsiunea de canalizare.

Câteva teste de bază în dezvoltarea fazei epidemiologice

În această secțiune, exemple lucrate de proceduri de calcul folosind R ilustrează:

> Testarea prezenței unui răspuns specific în tratamentele medicamentoase
> Testare pentru clasificarea încrucișată a datelor
> Estimarea parametrilor farmacocinetici

> **Exemplul 7.2:** Testarea prezenței bacteriilor după tratamente medicamentoase

Între 1999 și 2000, într-o investigație de sănătate în Teritoriul de Nord al Australiei, au fost întreprinse teste pentru a evalua efectele unui medicament asupra a 50 de copii cu antecedente de otită medie. Acești subiecți de caz au fost randomizați fie să ia medicamentul, fie să ia un placebo și, de asemenea, să primească încurajare activă pentru a se conforma cu regimul de medicamente. Prezența *Haemophilus influenzae* a fost verificată în săptămânile 0, 2, 4, 6 și 11. (Un total de 30 de verificări lipseau și nu sunt incluse în cadrul

[26] http://en.wikipedia.org/wiki/pro- și retro-studiu de cohortă.

de date.)

Acest cadru de date, numit bacterii , are 220 de rânduri și următoarele 6 coloane:

Y Prezența sau absența: un factor cu nivelurile n și y .
ap Activ/placebo: un factor cu nivelele a și p .
hilo Conformitate ridicată/scăzută: un factor cu niveluri hi și lo .
saptamana Numeric: saptamana de testare.
ID ID caz-subiect: un factor.
Trt Un factor cu niveluri placebo, droguri și droguri+, o înregistrare a ap și hilo .

Următorul segment de cod R poate fi utilizat pentru a efectua analiza:

```
> install.packages("MASS")
>   bibliotecă (MASS)
>   ls("pachet:MASĂ")
```

[1] „mănăstire”	**"acdeaths"**	**"addterm"**
[4] „Aids2”	**"Animale"**	**"anorexie"**
[7] „zonă”	**„ca.fracțiuni”**	"bacterii"
[10] „bandwidth.nrd”	**"bcv"**	**"beav1"**
[13] „beav2”	**"biopsie"**	**"naștere"**
[16] „Boston”	**"boxcox"**	**„varze**
[19] „caith"	**"Mașini93"**	**"pisici"**

[163] „waders” „whiteside” „width.SJ”
[166] „write.matrix” „wtloss”

```
>   install.packages(„supraviețuire")
>   bibliotecă (supraviețuire)
```

Se încarcă pachetul necesar: spline

```
>
>   date (bacterii)
>   atașează (bacterii)
```

Următoarele obiecte sunt mascate _de_ „.GlobalEnv”: y > bacterii

	da	ap	p	hilo	saptam	ID	trt
1				**salut**	**ana 0**	**X01**	**placebo**
2	**y**		**p**	**Hi**	**2**	**X01**	**placebo**
3	**y**		**p**	**Hi**	**4**	**X01**	**placebo**
4	**y**		**p**	**Hi**	**11**	**X01**	**placebo**
5	**y**		**o**	**Hi**	**0**	**X02**	**drog+**
219	**n**	**o**		**Hi**	**6**	**Z26**	**drog+**
220	**y**	**o**		**Hi**	**11**	**Z26**	**drog+**

>

```
>   contrasts(bacteria$trt) <- structure(contr.sdif(3),
+ dimnames = listă(NULL, c(„drog", „încurajează")))
>   ## analize cu efecte fixe
>   rezumat(glm(y ~ trt * săptămână, binom, date = bacterii))
```

Apel: glm(formula = y ~ trt * săptămână, familie = binom, date = bacterii)

Devianță Reziduuri:

Min	1Q	Median	3Q	Max
-2,2144	0,4245	0,5373	0,6750	1,0697

Coeficienți:

	Estimare Std.	Eroare	valoare	Pr(>|z|)
(Intercepta)	1,97548	0,30053	6.573	4.92e-11***
Trtdrug	-0,99848	0,69490	-1.437	0,15075
Trtencuraj	0,83865	0,73482	1.141	0,25374
Săptămână	-0,11814	0,04460	-2.649	0,00807 **
trtdrug:săptămână	-0,01722	0,10570	-0,163	0,87061
trtencuraj:saptama	-0,07043	0,10964	-0,642	0,52060 ---

Signif. coduri: 0 '*' 0,001 '**' 0,01 '*'0,05 '.' 0,1 " 1**

(Parametrul de dispersie pentru familia binomială considerată a fi 1) Devianța nulă: 217,38 pe 219 grade de libertate Devianță rezidual: 203,12 pe 214 grade de libertate AIC: 215,12 Număr de iterații Fisher Scoring: 4

```
>   rezumat(glm(y ~ trt + săptămână, binom, date = bacterii))
```

Apel: glm(formula = y ~ trt + săptămână, familie = binom, date = bacterii)

Reziduuri de devianta:

Min	1Q	Median	3Q	Max
-2,2899	0,3885	0,5400	0,7027	1,1077

Coeficienți:

	Estimare Std.	Eroare	valoare	Pr(>|z|)
(Intercepta)	1,96018	0,29705	6.599	4.15e-11 ***
trtdrug	-1,10667	0,42519	-2.603	0,00925 **
trtencuraj	0,45502	0,42766	1.064	0,28735
săptămână	-0,11577	0,04414	-2.623	0,00872 **--

Signif. coduri: 0 '*' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 " 1**

(Parametrul de dispersie pentru familia binomială considerată a fi 1) Devianța nulă: 217,38 pe 219 grade de libertate Devianță rezidual: 203,81 pe 216 grade de libertate AIC: 211,81 Număr de iterații Fisher Scoring: 4

```
>   rezumat(glm(y ~ trt + I(săptămâna > 2), binom, date = + bacterii))
```

Apel: glm(formula = y ~ trt + I(săptămâna > 2), familie = binom, date = bacterii)

Reziduuri de devianta:

Min	1Q	Median	3Q	Max
-2,4043	0,338	0,5754	0,6237	1,0051

Coeficienți:

	Estimare Std.	Eroare	valoarea_	Pr(>|z|)
(Intercepta)	2,2479	0,3560	6.315	2.71e-10 ***
trtdrug	-1,1187	0,4288	-2.609	0,00909 **
trtencuraj	0,4815	0,4330	1.112	0,26614
I (săptămâna > 2)	-1,2949	0,4104	-3.155	0,00160 **-- ADEVĂRAT

Signif. coduri: 0 '*' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 " 1**

(Parametrul de dispersie pentru familia binomială considerată a fi 1) Devianța nulă: 217,38 pe 219 grade de libertate Devianță reziduală: 199,18 pe 216 grade de libertate AIC: 207,18 Numărul de iterații Fisher Scoring: 5

```
> # analiză condiționată a efectelor aleatoare
> bibliotecă (supraviețuire)
> bacterii$Timp <- rep(1, nrow(bacterii))
> coxph(Surv(Timp, unclass(y)) ~ săptămână + strate(ID), + date = bacterii ,
  metoda = „exact")
```

Apel: coxph(formula = Surv(Time, unclass(y)) ~ săptămână + strate(ID), date = bacterii, method = "exact")

	Coef	exp(coef)	se(coef)	z	p
saptamana	-0,163	0,85	0,0547	-2,97	0,003

Testul raportului de probabilitate=9,85 pe 1 df, p=0,0017 n= 220, număr de evenimente= 177

```
>
> coxph(Surv(Timp, unclass(y)) ~ factor(săptămână) + strate(ID), + date =
  bacterii, metodă = „exact")
```

Apel: coxph(formula = Surv(Timp, unclass(y)) ~ factor(săptămână) + strate(ID), date = bacterii, metodă = „exact")

	coef	exp(coef)	se(coef)	z	p
factorul(săptăm	0,198	1.219	0,724	0,274	0,780
factorul(săptăm	-1.421	0,242	0,667	-2.131	0,033
factorul(săptăm	-1.661	0,190	0,682	-2.434	0,015
factorul(săptăm	-1.675	0,187	0,678	-2.471	0,013

Testul raportului de probabilitate=15,4 pe 4 df, p=0,00385 n= 220, număr de evenimente= 177

> coxph(Surv(Timp, unclass(y)) ~ I(săptămâna > 2) + strate(ID), + date = bacterii, metodă = „exact")

Apel: coxph(formula = Surv(Timp, unclass(y)) ~ I(săptămâna > 2) + strate(ID), date = bacterii, metodă = „exact")

	coef	exp(coef)	se(coef)	z	p
I(săptămâna > 2) ADEVĂRAT	**-1,67**	**0,188**	**0,482**	**-3,47**	**0,00053**

Testul raportului de probabilitate=15,2 pe 1 df, p=9,93e-05 n= 220, număr de evenimente= 177

>
> # Analiza PQL glmm
> bibliotecă (nlme)
> rezumat(glmmPQL(y ~ trt + I(săptămâna > 2), aleatoriu = ~ 1 | ID,
+ familie = binom, date = bacterii))

iterația 1
iterația 2
iterația 3
iterația 4
iterația 5
iterația 6
Modelul liniar cu efecte mixte se potrivește cu probabilitatea maximă

Date: bacterii

AIC	BIC	logLik
NA	NA	NA

Efecte aleatorii:
Formula: ~1 | ID

	(Interceptare)	Rezidual
StdDev:	**1,410637**	**0,7800511**

Funcția de variație:
Structura: greutati fixe
Formula: ~invwt
Efecte fixe: y ~ trt + I (săptămâna > 2)

	Valoare	Eroare standard	DF	valoarea t	valoarea p
(Intercepta)	**2,7447864**		**169**	**7,253294**	**0,0000**
trtdrug	**-1,2473553**	**0,6440635**	**47**	**-1,936696**	**0,0588**
trtencuraj	**0,4930279**	**0,6699339**	**47**	**0,735935**	**0,4654**
I (săptămâna > 2) ADEVĂRAT	**-1,6072570**	**0,3583379**	**169**	**-4,485311**	**0,0000**

Corelație:

	(Intr)	trtdrg	trtncr
trtdrug	**0,009**		
trtencuraj	**0,036**	**-0,518**	
I(saptamana > 2) ADEVARAT	**-0,710**	**0,047**	**-0,046**

Reziduuri standardizate în cadrul grupului:

Min	Q1	Med	Q3	Max
-5,1985361	0,1572336	0,3513075	0,4949482	1,7448845

Număr de observații: 220
Număr de grupuri: 50 >

Întrebări de revizuire pentru Secțiunea 7.2

1. (a) Numiți un domeniu din epidemiologie în care dezvoltarea procesului este de obicei realizată în mai multe faze.
   (b) Într-un astfel de proces, care sunt fazele obișnuite de dezvoltare? Descrieți pe scurt fiecare fază.
2. (a) Comparați un studiu de cohortă prospectiv cu un studiu de cohortă retrospectiv, dând un exemplu pentru fiecare.
   (b) Ce este farmacoepidemiologia? Dacă un studiu de cohortă este întreprins în această zonă a epidemiologiei, este prospectiv sau retrospectiv? De ce?
3. Descrieți și dați exemple de unele proceduri de testare care pot fi utilizate într-o fază epidemiologică tipică de dezvoltare?
4. Pentru dezvoltarea fazei epidemiologice, accesați pachetul CRAN MASS și selectați două funcții R pentru a testa prezența unui răspuns specific în tratamentul medicamentos . Demonstrați utilizarea lor prin exemple concrete adecvate.
5. Pentru dezvoltarea fazei epidemiologice, accesați pachetul CRAN PK și selectați două funcții R pentru estimarea parametrilor farmacocinetici în dezvoltarea fazei de proces. Demonstrați utilizarea lor prin exemple ilustrative adecvate.

Exerciții pentru Secțiunea 7.2

1. CRAN oferă o serie de pachete, concepute pentru analiza datelor de farmacocinetică, care sunt potrivite pentru sprijinirea dezvoltării fazei. Acestea sunt:
   (a) PK , PKfit , PKtools și nlmeODE. PK sprijină funcțiile farmacocinetice de bază și utilizează metode de analiză necompartimentală; celelalte trei suportă în principal metode de modelare.
   (b) Pachetele MASS și drc , care sprijină analiza răspunsurilor la doză.
   (c) Grila pachetului , pentru grafica spalier.
   Pentru o anumită sarcină în analiza dezvoltării fazei, se poate căuta o funcție R adecvată , adecvată sarcinii, pornind din mediul R după cum urmează:
      (i) Pornind de la opțiunea „Ajutor”, selectați „Pagina de pornire CRAN” din meniul derulant. Veți fi direcționat către pagina CRAN.
      (ii) Pe pagina CRAN, dintre opțiunile din partea stângă, selectați „Căutare”. Veți fi direcționat către pagina de căutare CRAN.
      (iii) Pe pagina de căutare CRAN, selectați „ R site search”; veți fi direcționat către pagina de căutare a site-ului CRAN R.
      (iv) În cele din urmă, introduceți subiectul de interes în caseta „Interogare” și selectați „Căutare”.

   (d) Dacă sarcina de dezvoltare a fazei la îndemână este de a testa o anumită

substanță biochimică în fiecare caz subiect după un anumit tratament medicamentos, utilizați pașii precedenți pentru a găsi o funcție R în CRAN care este potrivită pentru sarcină.

(e) Comparați rezultatul cu alegerea din Exemplul 7.2.

2. Pe lângă CRAN, există și alte platforme sursă pentru software de biostatistică care utilizează R ; două dintre ele sunt Bioconductor și PBSmodelling . Acestea pot fi accesate din următoarele surse:

(i) Pentru Bioconductor

- Începeți de pe site-ul de internet http://bioconductor.org
- Selectați „Instalare”.
- Selectați „Software” din tabelul cu versiuni de software .

(ii) Pentru PBSmodelling (Schnute, Couture-Beil, Haigh și Kronlund, 2011)

Pachetul R PBSmodelling conține software pentru a facilita proiectarea, testarea și operarea modelelor de computer . Inițialele *PBS* se referă la Stația Biologică Pacifică , un important laborator de pescuit de pe coasta Pacificului Canadei din Nanaimo, Columbia Britanică. Proiectat inițial pentru oamenii de știință din domeniul pescuitului, acest pachet are un potențial larg de aplicare în multe domenii științifice. PBSmodelling se concentrează în special pe instrumentele care facilitează construirea și editarea unei interfețe grafice personalizate adecvate pentru o anumită aplicație. (Pachetul este disponibil și de la CRAN.)

Repetați exercițiul 1, folosind o funcție R adecvată de la Bioconductor .

3. funcție R adecvată din PBSmodelling .
4. În pachetul CRAN PKtools (Blanchard, 2012) , care acceptă calcule pentru WinBUGS, NONMEM V, NLME și altele, funcția R diagtrplot() creează o diagramă trellis a concentrațiilor observate și a valorilor prezise în funcție de timp în funcție de subiect. Formatul său de utilizare este

diagtrplot(x,level,xvarlab,yvarlab,pages,...)

pentru care argumentele sunt

x Variabilă care identifică variabila de grupare.
nivel Nivelul modelului mixt (*p* —populație, *i* —individ).
xvarlab Etichetă pentru variabila *x* .
yvarlab Etichetă pentru variabila *y* .
pagini Numărul de pagini de imprimat; „1” afișează argumente suplimentare de pe prima pagină pentru a fi transmise funcțiilor de nivel inferior.

Următorul segment de cod R ilustrează o aplicație a acestui pachet:

```
> install.packages(„PKtools”)
> bibliotecă (PKtools)
Se încarcă pachetul necesar: zăbrele
Se încarcă pachetul necesar: nlme
Se încarcă pachetul necesar: R2HTML
Se încarcă pachetul necesar: xtable > ls(„pachet:PKtools”)
 [1]              „AICcomp”          „bugs”   „coVar.id”          „desc”
```

```
          „diagplot"
 [6]                „diagtrplot"   „HTMLtools"      „indEst"   „lonecpmt"
     „obvsprplot "
[11]                „paramEst"              „pk" „PKtools.AIC"  „residplot"
          „RunNLME"
[16]                „RunNM" „RunWB" „sonecpmt"  „tex" „         trplot "
> bibliotecă (nlme)
> date(Theoph)
>   Theoph <- Theoph[Theoph$Time!=0,]
>   id <- as.numeric(as.character(Theoph$Subiect))
>   doza <- Theoph$Doza
>   timp <- Theoph$Timp
>   conc <- rotund(sqrt(Theoph$conc),4)
>   Theo <- data.frame(cbind(id,doza,time,conc))
> nume(Theo) <- c("id","doza","timp","conc")
>   wt.v <- Theoph$Wt
>   date <- list(pkvar = Theo, cov = wt.v)
> nameData <- list(covnames = c("wt"),
+                      yvarlab = "Sqrt(Theop. Conc.)   (mg/L)",
+                      xvarlab = „Timp de la doză (ore)",
+                      reparams = c("Cl"),
+                      parametri = c("Ka","V", "Cl"),
+                      tparams = c("log(Ka)","log(V)","log(CL)"))
>   model.def <- list(fixed.model=lKa+lV+lCl~1,
>                      random.model=lCl~1,
+                      start.lst=c(lKa=.3,lV=-.6,lCl=-3),
+                      form=conc~sonecpmt(doză, timp,
+                              lV, lKa, lCl),
+                      control=nlmeControl(returnObject=FALSE))
>   MM<-RunNLME(inputStructure=model.def,data=data,
+                      nameData=nameData)
>   diagtrplot(x=MM,level="p", xvarlab=nameData$xvarlab,
+              yvarlab=nameData$xvarlab, pages=1)
>   # Ieşire: Figura 7.10 .
```

(a) Explicați funcția fiecărei linii a segmentului de cod R pentru acest calcul.

(b) Reluați acest segment de cod în mediul R.

(c) Comparați și comentați codul R : valorile prezise versus valorile observate ale acestui set de date.

5. Calcule R pentru farmacocinetica indometacinei.

Pachetul CRAN nlmeODE (Tornoe, 2012) (modelare neliniară cu efecte mixte folosind ecuații diferențiale obișnuite) combină de fapt două pachete: nlme și Ode solve , pentru modelarea cu efecte mixte folosind ecuații diferențiale. Următorul segment de cod R modelează farmacocinetica indometacinei folosind acest pachet CRAN:

```
> instalați pachete ("nlmeODE")
pachetul odesolve despachetat cu succes și sumele MD5 verificate
```

pachetul nlmeODE a fost despachetat cu succes și sumele MD5 au fost verificate

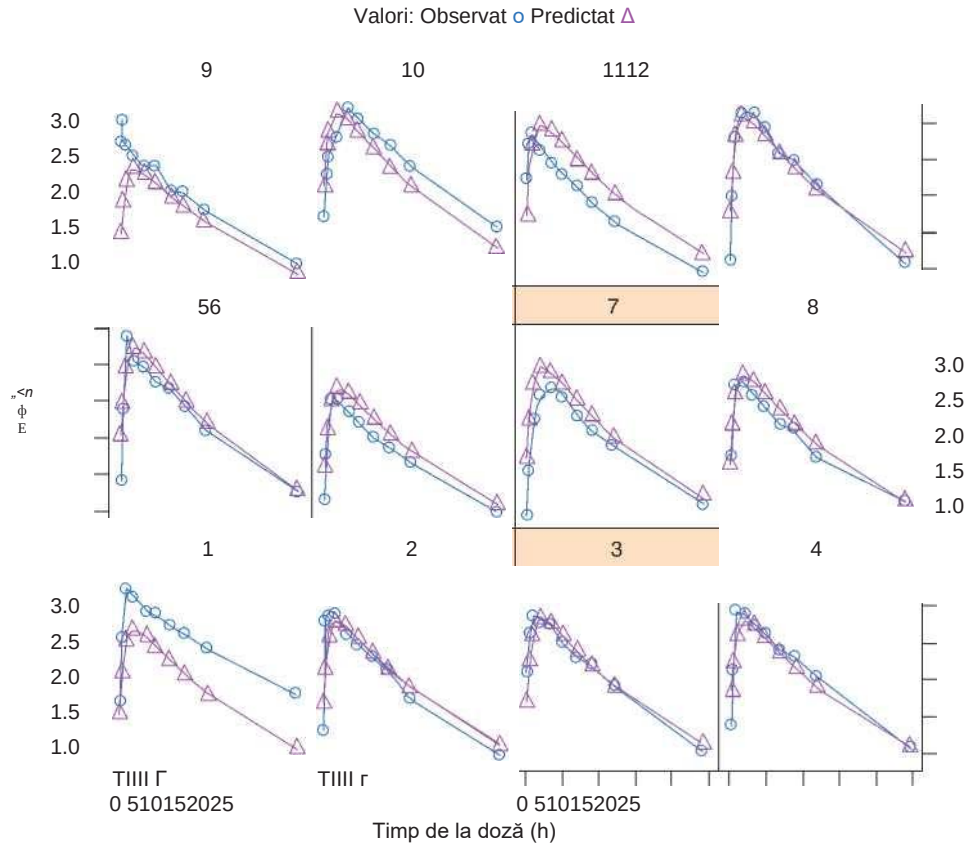


FIGURA 7.10 Utilizarea pachetului CRAN PKtools.

```
> bibliotecă (nlmeODE)
Se încarcă pachetul necesar: odesolve
odesolve este depreciat! Utilizați soluțiile din deSolve.
odesolve va fi eliminat din CRAN până la sfârșitul anului 2012.
Se încarcă pachetul necesar: nlme
Se încarcă pachetul necesar: zăbrele

> ls("pachet:nlmeODE")
[1]  "nlmeODE"
>

###################################
### Farmacocinetica indometacinei ###
###################################

> date (Indometh)
> Indometh # Ieșirea cadrului de date Indometh pentru inspecție:
```

Date grupate: conc ~ timp | Subiect

	Timp	subiect	Conc
1	**1**	**0,25**	**1,50**
2	**1**	**0,50**	**0,94**
3	**1**	**0,75**	**0,78**
4	**1**	**1,00**	**0,48**
5	**1**	**1,25**	**0,37**
64	**6**	**5,00**	**0,13**
65	**6**	**6,00**	**0,10**
66	**6**	**8,00**	**0,09**

```
>
> TwoComp <- list(DiffEq=listă(
+ dy1dt = ~ -(k12+k10)*y1+k21*y2,
+ dy2dt = ~ -k21*y2 + k12*y1),
+ ObsEq=list(
+ c1 = ~ y1,
+ c2 = ~ 0),
+ State=c("y1","y2"),
+ Parms=c("k12","k21","k10","start"),
+ Init=list("start",0))
>   IndomethModel <- nlmeODE(TwoComp,Indometh)
>   Indometh.nlme <- nlme(conc ~
+ IndomethModel(k12,k21,k10,start,time,subject),
+ date = Indometh, fix=k12+k21+k10+start~1,
+ aleatoriu = pdDiag(start+k12+k10~1),
+ start=c(k12=-0,05,k21=-0,15,k10=-0,10,start=0,70),
+ control=listă(msVerbose=TRUE),
+ verbose=TRUE)
```

0: -12,615270: -0,785113 -1,41611 -0,440966
1: -12,615270: -0,785095 -1,41627 -0,440425
2: -12,615270: -0,785095 -1,41627 -0,440425

**** Iterația 1**
Pas LME: Loglik: 57.22393, iterații nlm: 2 parametri reStruct:

Subiect1	**Subiect2**	**Subiect3**
-0,7850948	**-1,4162729**	**-0,4404247**

Pasul PNLS: RSS = 0,4128427

efecte fixe:	**-0,0331751**	**-0,115947**	**-0,111574**	**-0,717857**

iterații: 7
Convergență:

reStruct fix
0,5071548 0,2703262

0: -11,974429: -0,766570 -1,43832 -0,347128
1: -11,974429: -0,766340 -1,43872 -0,345577
2: -11,974429: -0,766497 -1,43869 -0,345660
3: -11,974430: -0,766594 -1,43863 -0,346103
4: -11,974430: -0,766487 -1,43861 -0,346124
5: -11,974430: -0,766513 -1,43851 -0,346167
6: -11,974430: -0,766544 -1,43859 -0,346197
7: -11,974430: -0,766527 -1,43858 -0,346173
8: -11,974430: -0,766529 -1,43857 -0,346176

**** Iterația 2**
Pas LME: Loglik: 56.58309, iterații nlm: 8 parametri reStruct:

Subiect1 Subiect2 Subiect3
-0,7665286 -1,4385742 -0,3461758

Pasul PNLS: RSS = 0,4128658
efecte fixe:-0,0331751 -0,115947 -0,111574 0,717857 iterații: 1
Convergență:

ReStruct fix
0,000000000 0,002752387

0: -11,974430: -0,766529 -1,43857 -0,346176
1: -11,974430: -0,766529 -1,43857 -0,346176

**** Iterația 3**
Pas LME: Loglik: 56.58309, iterații nlm: 1 parametri reStruct:

Subiect1 Subiect2 Subiect3
-0,7665286 -1,4385745 -0,3461757

Pasul PNLS: RSS = 0,4128658 efecte fixe:-0,0331751 -0,115947 -0,111574 0,717857
iterații: 1
Convergență:

reStruct fix
0,00000e+00 3,03061e-08

```
> plot(augPred(Indometh.nlme,level=0:1))
> # Ieşire: Figura 7.11 .
```

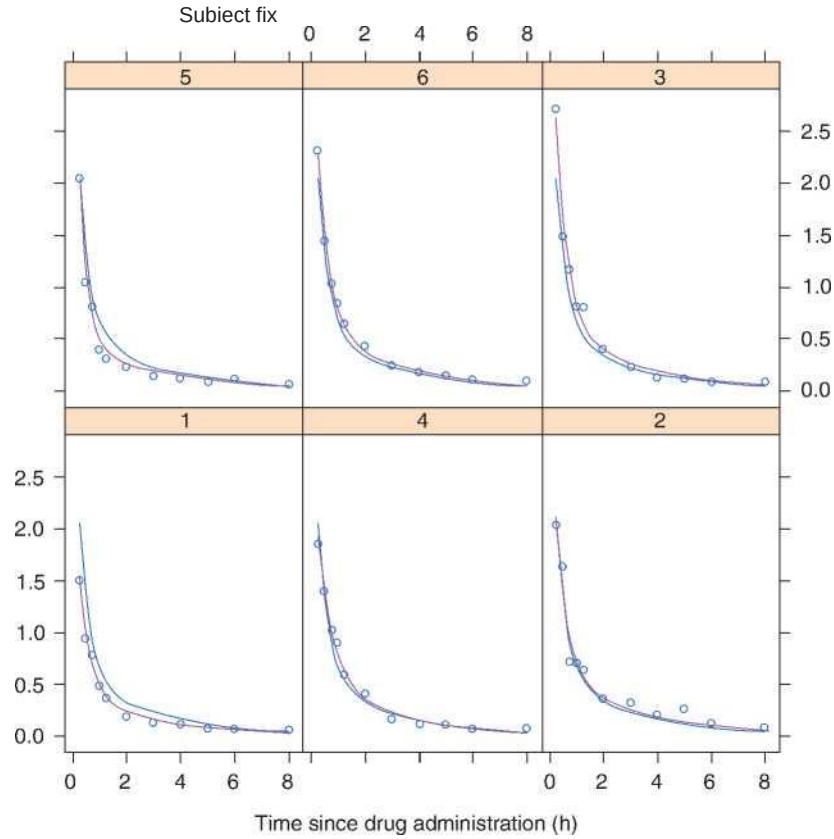


FIGURA 7.11 Farmacocinetica indometacinei prin nlmeODE.

(a) Explicați funcția fiecărei linii a segmentului de cod R pentru acest calcul.
(b) Rulați din nou acest segment de cod în mediul R , înlocuind ultima comandă a segmentului de cod cu următoarele:

```
> plot(augPred(Indometh.nlme))
```

și apoi rulați din nou întregul segment de cod. Comentează noua parcelă rezultată.

7.3 CONFUNDARE ÎN ANALIZA SUPRAVIEȚIEIULUI

În analiza supraviețuirii (vezi Capitolul 5 , Secțiunea 5.2), precum și în majoritatea investigațiilor epidemiologice, trebuie să ne preocupăm cum să caracterizam asocierea dintre două variabile (de obicei, un rezultat ca răspuns la o expunere) (Children's Spitalul Mercy, 2000) . Un exemplu tipic este dorința de a testa dacă nivelul total de colesterol (rezultatul) al unui subiect de caz este asociat cu consumul de alcool (expunerea). Cu toate acestea, datele cercetărilor anterioare par să indice că statutul de fumat al subiectului cazului are un efect profund asupra acestei relații. În această situație, epidemiologii pot considera statutul de fumat al cazului subiect ca fiind un potențial - confuz **în** această relație, deoarece statutul de fumat al subiectului cazului.

tinde să fie strâns asociat cu consumul intens de alcool și este, de asemenea, legat de nivelul de colesterol în rândul persoanelor care *nu sunt* consumatori mari de alcool.

Un *factor de confuzie* este definit ca o variabilă epidemiologică care este:

- asociat cu variabila de expunere,
- asociat, *independent , cu* variabila rezultat și
- nu într -o cale cauzală directă între expunere și boală.

Acest concept este ilustrat în Figura 7.12 .

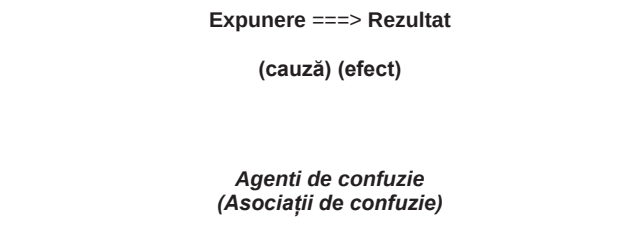


FIGURA 7.12 O vedere schematică a confuziei în epidemiologie.

În această reprezentare:

- Săgeata ruptă cu un singur cap denotă o cauzalitate epidemiologică.
- Săgețile solide cu două capete denotă asocieri cu factori de confuzie.

Scopul analizei este de a determina cauzalitatea epidemiologică în timp ce se ajustează pentru efectele *tuturor* asociațiilor de confuzie.

Abordări biostatistice pentru controlul confuziei

În investigațiile epidemiologice, confuzia poate fi controlată prin utilizarea controalelor experimentale sau a controalelor experimentale și analitice combinate.

CONTROALE EXPERIMENTALE DE CONFUNDARE

Modalitățile de modificare a unui design de studiu epidemiologic pentru a exclude sau controla în mod activ variabilele confundatoare includ [27]:

- **Studii caz-control** . Această abordare atribuie factori de confuzie ambelor grupuri – cazuri și controale – în mod egal. De exemplu, atunci când se studiază cauza infarctului miocardic, vârsta subiecților cazului poate fi considerată o variabilă probabilă de confuzie. Să presupunem că investigatorul încearcă să potrivească fiecare pacient cu infarct de 59 de ani cu o persoană „control" sănătoasă de 59 de ani. (În studiile caz-control, variabilele potrivite cel mai adesea sunt vârsta și sexul.)

  Dezavantaj: studiile caz-control sunt fezabile numai atunci când este ușor de găsit controale; adică persoane pentru care toți factorii de confuzie potențiali cunoscuți sunt la fel ca și pentru subiectul cazului. De exemplu, un investigator al

[27] Derutant. Preluat de la http://en.wikipedia.org/wiki/Confounding

unui studiu caz-control

încearcă să găsească cauza unei anumite boli la o persoană care are (a) 39 de ani, (b) chinezo-american, (c) din Dakota de Nord, (d) un jucător de tenis dedicat, (e) un vegan sau un vegetarian și (f) lucrează ca pompier. Un control teoretic perfect ar fi o persoană care, pe lângă faptul că *nu* are boala investigată , se potrivește și cu toate aceste șase caracteristici și nu are boli pe care subiectul cazului să nu le aibă. Evident, găsirea unui astfel de control (darămite suficient pentru a face un grup de control) ar fi o adevărată provocare!

- **Studii de cohortă** . Această abordare face potrivirea apropiată mai practică prin controlul admiterii anumitor grupe de vârstă sau a unui anumit gen în populația studiată , creând astfel o cohortă de persoane care împărtășesc caracteristici similare și făcând comparabile toate cohortele în ceea ce privește posibila variabilă de confuzie. De exemplu, dacă se consideră că vârsta și sexul sunt factori de confuzie, doar bărbații cu vârsta cuprinsă între 30 și 40 de ani ar fi admiși într-un studiu de cohortă care intenționează să evalueze riscul de infarct miocardic în cohorte care sunt fie activi fizic, fie inactivi.

  Dezavantaj: în studiile de cohortă, supraexcluderea datelor de intrare poate determina biostatisticienii să definească prea restrâns setul de persoane situate în mod similar pentru care pot pretinde că studiul este util; alte persoane cărora li se aplică de fapt relația de cauzalitate pot pierde oportunitatea de a beneficia de recomandările studiului. În mod similar, suprastratificarea datelor de intrare în cadrul unui studiu poate reduce dimensiunea eșantionului dintr-un strat dat până la punctul în care generalizările extrase din observațiile membrilor acelui strat pot să nu fie semnificative statistic.
- **Dublu orb** . Această abordare ascunde membrii grupului de experiment a participanților de populația de studiu *și* de observatori. Când participanții sunt împiedicați să știe dacă primesc sau nu tratament, orice efecte placebo probabile ar trebui să fie aceleași pentru grupul de control și pentru grupul de tratament. Atunci când observatorii sunt împiedicați să cunoască apartenența participanților, nu ar trebui să existe nicio părtinire din partea cercetătorilor care tratează grupurile în mod diferit sau de a interpreta rezultatele în mod diferit.
- **Studiu randomizat controlat (RCT)** . În această abordare, populația de studiu este împărțită aleatoriu pentru a reduce șansele de auto-selecție de către participanți sau părtinire de către proiectanții studiului. Înainte de începerea experimentului, testatorii vor atribui membrii grupului de participanți grupurilor lor (control, intervenție și paralel), folosind un proces de randomizare , cum ar fi utilizarea unui generator de numere aleatorii. De exemplu, într-un studiu privind efectele exercițiilor fizice, concluziile ar fi mai puțin valide dacă participanților li s-ar oferi posibilitatea de a alege dacă doresc să aparțină grupului de control (care nu ar exercita) sau grupului de intervenție (care ar avea nevoie de parte într-un program de exerciții). Studiul ar capta apoi și alte variabile în afară de exerciții fizice, cum ar fi nivelurile de sănătate pre-experiment și motivația de a adopta activități sănătoase. Experimentatorul poate alege subiecte de caz care au mai multe șanse să arate rezultatele pe care studiul speră să le găsească sau pot interpreta rezultatele subiective (mai energice, atitudine mai pozitivă) într-un mod favorabil preferințelor experimentatorului.
- **Stratificare** . Pentru a explica stratificarea, să revenim la exemplul anterior în care investigația examinează cauzele infarctului miocardic. Se crede că activitatea fizică

reduce apariția infarctului miocardic și se presupune că vârsta este un posibil factor de confuzie. Astfel, datele eșantionate sunt
apoi stratificat pe grupe de vârstă; aceasta înseamnă că asocierea dintre activitate și infarct va fi analizată pe grupe de vârstă. Dacă diferitele grupuri de vârstă (sau straturi de vârstă) generează **rapoarte de risc foarte diferite** , vârsta trebuie privită ca o variabilă de confuzie. Unele instrumente statistice, printre care metodele Mantel–Haenszel (M–H), pot lua în considerare stratificarea seturilor de date (vezi Exemplul 7.3) .

Dintre aceste cinci abordări pentru excluderea sau controlul confuziei, primele patru au fost discutate anterior (vezi Secțiunile 6.1 și 6.2 din Capitolul 6 și Secțiunea 7.1). Ultima abordare, stratificarea, este discutată mai târziu în această secțiune.

CONTROL COMBINAT EXPERIMENTAL ȘI ANALIT AL CONFUNDING (FITZMAURICE, 2004)[28]

În general, două abordări sunt utilizate pentru a constrânge și a reduce confuzia și pentru a ajusta analiza datelor pentru efectele confuziei: analiza prin stratificare sau utilizarea modelării regresiei. Ambele abordări sunt discutate și ilustrate în această subsecțiune.

AJUSTAREA CONFUNDEI PRIN STRATIFICARE. Stratificarea este o modalitate stabilită și eficientă de ajustare pentru confuzie în analiza datelor experimentale cauza-efect (expunere și rezultat). Luați în considerare următorul set de date ipotetice dintr-o investigație privind asocierea dintre administrarea de suplimente de vitamina E și riscul de apariție a bolilor coronariene, prezentat în Figura 7.13 .

		CHD	
		Prezent	**Absent**
Vitamina E	**Da:**	60	601
	Supliment **nr:**	78	461

FIGURA 7.13 Asocierea dintre suplimentarea cu vitamina E și riscul de CHD (cadru de date ipotetic).

Acest set de date a fost preluat de la 1200 de subiecți cu vârste cuprinse între 40 și 45 de ani, care nu erau lipsiți de CHD la începutul studiului; au fost urmăriți timp de 20 de ani. Pe parcursul perioadei de urmărire de 20 de ani, cei care au dezvoltat CHD au fost identificați . Subiecții cazului au fost clasificați în funcție de utilizarea de suplimente de vitamina E (variabila „expunere").

Analiza datelor prezentate în Figura 7.13 indică faptul că:

O1 = șansele de CHD pentru cei care au folosit suplimente de vitamina E sunt 60/601 = 0,100 (rotunjit); și

O2 = șansele de CHD pentru cei care *nu au* folosit suplimente de vitamina E sunt 78/461

[28] Tratarea confuziei în analiză. Preluat de pe www.iarc.fr/en/publications/pdfs- online/epi/cancerepi/CancerEpi-14.pdf

= 0,169 (rotunjit).

Această asociere aparentă poate fi exprimată în termeni de raportul de cote (OR):

$$SAU = O1/O2 = 0,100/0,169 = 0,592$$

Acest OR estimat poate implica faptul că șansele de a dezvolta CHD sunt reduse cu aproape 50%; adică, șansele de a dezvolta CHD sunt aproape înjumătățite prin administrarea de suplimente cu vitamina E. Cu toate acestea, o problemă potențială cu această analiză și concluzie este că *aceste două grupuri pot diferi în alte moduri decât utilizarea suplimentelor de vitamina* E.

Să presupunem că am prefera să luăm în considerare factorii de risc asociați. Apoi

R1 = Riscul de CHD pentru cei care au folosit suplimente de vitamina E este 60/(60 + 601) = 60/661 = 0,091

și

R2 = Riscul pentru cei care nu au folosit suplimente de vitamina E este 78/(78 + 461) = 78/539 = 0,145

dând un raport de risc de

$$RR = R1/R2 = 0,091/0,145 = 0,628$$

și implicând un risc substanțial mai mic, în acord substanțial cu rezultatul anterior.

Ca exemplu de confuzie, subiecții de caz care au folosit suplimente de vitamina E pot fi mult mai puțin susceptibili de a fuma, caz în care trebuie să existe o anumită îngrijorare că asocierea observată în Figura 7.13 se datorează, cel puțin parțial, **confuziei .** efectele fumatului (fumatul este un factor de risc major și bine stabilit pentru CHD). Pentru a face față acestei posibilități, se pot **stratifica** (clasifica în grupuri cu caracteristici similare și definitive) datele din Figura 7.13 în funcție de faptul dacă un subiect de caz este fumător. Vezi Figura 7.14 .

Fumatori:

		CHD	
		Prezent	Absent
Vitamina E Supliment nr:	Da:	13	48
		59	240

Nefumători:

		CHD	
		Prezent	Absent
Vitamina E Supliment nr:	Da:	47	553
		19	221

FIGURA 7.14 Rezultatele stratificate după istoricul de fumat: asociere între suplimentarea cu vitamina

E și riscul de CHD (cadru de date ipotetic).

Din aceste date stratificate, șansele de a ceda la CHD în rândul clasei care utilizează suplimente cu vitamina E, în raport cu clasa care nu o folosesc, pot fi acum calculate separat pentru fumători și pentru nefumători.

Pentru fumători, OR estimat este

$$\text{SAU (fumători)} = (13/48)/(59/240) = 0{,}271/0{,}246 = 1{,}097$$

Pentru nefumători, OR estimat este

$$\text{SAU (nefumători)} = (47/533)/(19/221) = 0{,}088/0{,}086 = 1{,}023$$

Aceste rezultate par să indice că, după controlul efectelor fumatului , există puține dovezi sau deloc ale unei asocieri între suplimentarea cu vitamina E și riscul de a suferi de CHD.

Nu ar trebui să scape de atenție că acest rezultat aparent paradoxal se datorează faptului că grupul de suplimente cu vitamina E conține semnificativ mai puțini fumători: 360 (= 13 + 48 + 59 + 240), ceea ce reprezintă doar 30% din totalul populației de caz-subiect. ulare de 1.200. Astfel, 70% din totalul populației testate sunt nefumători, care au avut un risc mai mic de a ceda la CHD.

Analiza suplimentară de mai sus, bazată pe datele stratificate din Figura 7.14 , demonstrează o tehnică eficientă pentru ajustarea efectelor confuziei în analiza inițială. În această abordare, cunoscută sub numele de **stratificare** , efectul de confuzie este controlat prin stratificarea nivelurilor potențialului factor de confuzie (care, în acest exemplu, este istoricul de fumat al subiecților cazului). În special, confuzia este - controlată atunci când asocierea de interes este inclusă în grupuri distincte, sau *straturi* , alcătuite din subiecți care sunt altfel omogene în raport cu variabila de confuzie.

În acest moment, ar trebui să fie evident că un factor care poate avea o influență importantă asupra analizei este dimensiunea eșantionului, precum și dimensiunile eșantionului relative ale diferitelor straturi din cadrul populației totale. Această problemă poate fi abordată folosind metoda MH de stratificare.

METODA MH DE STRATIFICARE. Metoda MH de stratificare calculează o măsură sumară a asocierii, luând o medie ponderată a estimărilor specifice stratului, folosind ponderi proporționale cu dimensiunea eșantionului din fiecare strat. Aceste ponderi sunt invers proporționale cu variațiile estimărilor specifice stratului OR, dând astfel o pondere mai mare estimărilor mai precise.

Luați în considerare estimarea comună a OR. Să existe *K* niveluri de stratificare, fiecare nivel luând o variabilă de stratificare separată: K = 2 pentru setul de date din Figura 7.14 . Datele pot fi apoi rezumate în termeni de tabele de contingență *K* 2x2. În fiecare tabel, cele patru celule interne conțin numărătoare de frecvență ale numărului de subiecți de caz care au o combinație particulară a celor două variabile. Figura 7.15 prezintă un astfel de tabel generic în care rândurile corespund unei expuneri (în acest caz, suplimentarea cu vitamina E), iar coloanele corespund stării bolii (în acest caz, prezența sau absența CHD).

	BOALA	
	Prezent	Absent
Da:	**a** $_k$	**b** $_k$

Expunere (cauza)

Nu: **c** $_k$ **d** $_k$

FIGURA 7.15 Tabel de contingență 2x2 specific stratului pentru analiza MH.

Formula MH pentru estimarea combinată a OR, derivată dintr-o serie de tabele de contingență 2x2, este

$$\text{SAU}_{MH} = \{\sum_{k=1}^{K} (a_k d_k)/n_k\}/\{\sum_{k=1}^{K} (b_k c_k)/n_k\} \qquad (7.1)$$

unde n_k este numărul total de observații din k -lea tabel ($n_k = a_k + b_k + c_k + d_k$) și însumarea din ecuația (7.1) este preluată peste K nivelurile variabilei de stratificare . De exemplu, estimarea combinată MH a OR bazată pe datele din cele două tabele de contingență din Figura 7.14 este calculată după cum urmează:

$$\begin{aligned}\text{ORMH} &= \{(13 \times 240)/360 + (47 \times 221)/840\}/\{(48 \times 59)/360 + (553 \times 19)/840\}\\ &= (7{,}333 + 12{,}365)/(7{,}867 + 12{,}508)\\ &= 19{,}698/20{,}375\\ &= 0{,}967\end{aligned}$$

După cum era de așteptat pentru aceste date, estimarea combinată a OR este de aproximativ 1,0, indicând că *nu există dovezi de asociere* între suplimentarea cu vitamina E și riscul de a ceda la CHD, după controlul efectelor istoricului de fumat.

Mai mult, poate fi calculat un interval de încredere (CI) pentru OR ajustat. Limitele IC de 95% pentru OR ajustat sunt în intervalul [0,14, 6,33], care include valoarea nulă de 1,0 pentru OR. Acest IC de aproximativ 95% pentru ln OR_{MH} este estimat:

$$\begin{aligned}95\%\ \text{CI}\ (\ln \text{SAU}_{MH}) &= (\ln 0{,}95) \pm 1{,}96 \times 0{,}967\\ &= -0{,}05 \pm 1{,}96 \times 0{,}967\\ &= -0{,}05 \pm 1{,}895\\ &= -1{,}945 \text{ până la } 1{,}845\end{aligned}$$

Un CI de aproximativ 95% pentru OR_{MH} poate fi obținut prin luarea de antilogaritmi:

$$\begin{aligned}95\%\ \text{CI}\ [\text{OR}_{MH}] &= [e^{-1{,}945}, e^{1{,}845}]\\ &= [0{,}14, 6{,}33]\end{aligned}$$

Ultimul rezultat poate fi obținut prin calcule folosind următorul segment de cod R :

```
> exp(-1,945)
[1] 0,1429872
> exp(1.845)
[1]  6,3281
```

Observatii:

1. Atunci când se evaluează dovezile pentru o asociere între două variabile, trebuie să se ia întotdeauna în considerare efectul factorilor de confuzie (măsurați sau nu). Dacă nu este ajustată, confuzia poate duce la concluzii inacceptabile: un efect cauzal poate fi sugerat acolo unde nu există, sau un efect adevărat poate fi ascuns. Fiecare concluzie eronată poate duce la o interpretare sau un diagnostic incorect.
2. Din cauza acestor posibilități de eroare, confuzia pentru care nu poate fi controlată în proiectarea studiului ar trebui ajustată în analiză. Procedura MH, în special atunci când este efectuată de computer, oferă o medie ponderată a estimărilor de asociere nefondate, specifice stratului.
3. Atunci când există mulți factori potențiali de confuzie, care rezultă în straturi cu prea puține subiecți pentru a face comparații semnificative, stratificarea poate fi mult mai puțin atractivă.
4. Acest lucru conduce la o abordare alternativă pentru ajustarea pentru confuzie în analiză : examinați efectul expunerii într-un **model de regresie** pentru dependența rezultatului bolii de expunerea de interes și orice potențiali factori de confuzie. Această abordare este discutată în subsecțiunea următoare.

AJUSTAREA CONFUNDĂRII PRIN MODELAREA REGRESIEI (McNamee, 2005) . Ca o introducere în analiza de regresie, luăm în considerare cazul a două variabile măsurate pe scale continue, cu scopul de a investiga influența unei variabile asupra alteia. Aici, analiza de regresie este utilizată pentru a începe aplicarea controlului factorilor de confuzie.

■ **Exemplul 7.3:** Ajustarea confuziei prin modelarea regresiei

Pasul 1: Începând cu un model de bază simplu

Într-o anumită populație, anchetatorii iau în considerare 160 de subiecți nefumători. Variabile sunt vârsta și scăderea odată cu vârsta a volumului expirator forțat (FEV) în 1 secundă (FEV). Datele privind ambele variabile, vârsta și VEMS, au fost adunate dintr-un grafic încrucișat al datelor, așa cum se arată în Figura 7.16 .

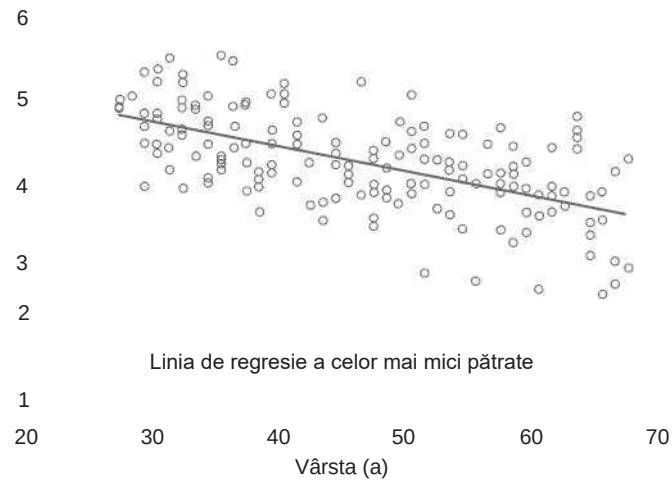


FIGURA 7.16 Relația dintre VEMS și vârstă la 160 de bărbați nefumători.

Pe baza acestor date, se poate propune un model al relației în populație, unde modelul este specificat printr-o formă model sau ecuații model. Dacă se consideră o relație liniară, modelul ar avea forma generală:

$$\text{Model 1: VEMS} = a + b\ (\text{Vârsta}) + r \qquad (7,2)$$

Cele trei cantități necunoscute din acest model — *a* , *b* și *r* — pot fi estimate în analiză.

Amintiți-vă că în geometria analitică sau de coordonate, ecuația generală pentru o dreaptă în planul X–Y este

$$y = m_1 x + b_1 \qquad (7,3)$$

unde m_1 este panta sau gradientul dreptei, iar b_1 este ordonata (valoarea y) a interceptului dreptei cu axa *y* : *b* = *y* (la *x* = 0).

Modelul în care *r* este ignorat (prin stabilirea *lui r* = 0) este o descriere a relației dintre „vârstă" și „FEV" mediu în rândul persoanelor de o anumită vârstă. *r* este o componentă aleatorie care se presupune că variază de la o persoană la alta. Includerea acestui termen în model permite recunoașterea faptului că persoanele de aceeași vârstă nu sunt toți la fel și, prin urmare, valorile lor individuale ale VEMS vor varia în jurul mediei pentru acea vârstă. Variația aleatoare este imprevizibilă, dar poate fi descrisă printr-o distribuție statistică. Cu variabile continue, cum ar fi VEM, se presupune adesea că componenta aleatorie are o distribuție normală cu o medie de zero.

Metodele statistice pot fi utilizate apoi pentru a estima coeficienții de regresie, *a* și *b* și s(*r*), abaterea standard a *lui r* , dintr-un eșantion de date. Astfel, în metoda de estimare a celor mai mici pătrate, rațiunea este de a alege valori pentru *a* și *b* care minimizează s(*r*) în setul de date. Aplicarea acestei metode la datele din Figura 7.16 oferă estimări pentru:

a = 5,58 L
b = - 0,03 L/an de vârstă și
s(*r*) = 0,46 L

Prin urmare, modelul „montat" este

$$\text{VEMS} = 5,58 - 0,039\ (\text{Vârsta}) + r \qquad (7,4)$$

Tabelul 7.2 prezintă o ieșire tipică dintr-un program de calculator biostatistic care se potrivește cu modelul 1 la date, inclusiv 95% CI pentru *a* și *b* și valorile *p* pentru testele de semnificație. În fiecare test, ipoteza nulă este că valoarea adevărată a coeficientului este zero. Astfel, dacă *b* = 0, atunci vârsta nu ar avea niciun efect asupra VEMS. În acest caz, atât testul de semnificație, cât și IC de 95% sugerează că *b* < 0.

TABELUL 7.2 Ieșirea tipică computerizată de la modelul de adaptare 1 la estimarea datelor modelului 1
Coeficienți (din datele din figura 7.16)

FEV	COEFICIENT	EROARE STD	T STATISTIC	PROBABILITATE	95% CI
Vârstă	*b* = - 0,0301	0,0032	- 9,52	< 0,001	- 0,0363 până la - 0,0238
Constant	*a* = 5,5803	0,1440	38,75	< 0,001	5,2960 până la 5,8647

Eroare pătratică medie, s(*r*) = 0,464 L.

Observatii:

1. În Exemplul 7.3 , Modelul 1 este un model liniar, care presupune că VEMS mediu scade cu o cantitate fixă, -0,0301 litri (30,1 mililitri), pe an în vârstă.
2. Rețineți că liniaritatea a fost *presupusă* , nu dovedită și că analiza biostatistică a estimat pur și simplu coeficienții unui model liniar presupus.
3. Sunt disponibile metode suplimentare de diagnosticare cu regresie postestimare (cum ar fi analiza reziduurilor și diferite măsuri cantitative pentru identificarea observațiilor influente) pentru a evalua bunătatea potrivirii. Astfel, se poate duce modelarea la un nivel superior, așa cum este ilustrat în următorul exemplu lucrat.

Pasul 2: Extinderea modelului de bază cu variabile dependente suplimentare: ajustarea suplimentară a confuziei prin modelarea regresiei

Se crede că, pentru fiecare subiect de caz, înălțimea corpului și istoricul de fumat (numărul de țigări fumate pe zi) pot afecta VEMS. Modificarea modelului de regresie 1, ecuația (7.2) , pentru a include aceste variabile suplimentare și *presupunând că fiecare variabilă dependentă are o relație liniară cu VEMS* și *că efectul comun al celor trei variabile (factori) dependente este suma efectelor lor separate* , se poate scrie:

$$\text{Model 2: VEMS} = a + b\,(\text{Vârsta}) + c\,(\text{Înălțime}) + d\,(\text{Țigări}) + r \qquad (7,5)$$

Cu acest model și același set de date, o analiză statistică standard ar produce estimări ale coeficienților a , b , c și d și s(r); CI de 95%; iar testul ipotezei nule pentru fiecare coeficient.

Se poate postula că, deoarece VEMS măsoară volumele, ar trebui de așteptat ca VEMS să varieze ca cubul *de* înălțime (Înălțime) ^{3} , iar acest lucru ar trebui luat în considerare pentru Modelul 2, pentru a da Modelul 3:

$$\text{Modelul 3: VEMS} = a + b\,(\text{Vârsta}) + c\,(\text{Înălțimea})^3 + d\,(\text{Țigări}) + e\,(\text{Grup}) + r \qquad (7,6)$$

Cu toate acestea, analiza de regresie poate ajuta ce formă de model este mai potrivită, liniară sau cubică.

De asemenea, se poate lua în considerare dacă amploarea efectului fumatului ar trebui să varieze cu vârsta; acest efect de interacțiune este cunoscut sub numele de **modificare a efectului** . În acest scop, s-ar putea postula un termen de produs suplimentar care combină vârsta și țigările. Acest lucru ar modifica Modelul 2 pentru a oferi Modelul 4:

$$\text{Modelul 4: VEMS} = a + b\,(\text{Vârsta}) + c\,(\text{Înălțimea})^3 + d\,(\text{țigări}) + e\,(\text{Vârsta})(\text{Țigări}) + f\,(\text{Grup}) + r \qquad (7,7)$$

Din nou, aceasta va fi urmată de analiză pentru evaluarea coeficienților: a , …, f , r . În mod clar, trebuie să decideți cât de departe ar trebui să se ducă acest proces.

O altă metodă generală disponibilă este modelul aditiv generalizat (GAM) (Hastie & Tibshirani, 1990 ; Wood, 2006) , care ia următoarea formă:

$$\text{Model 5: GAM: g(E[FEV])} = a_0 + f_1(\text{Vârsta}) + f_2(\text{Înălțime}) + f_3(\text{Țigări}) + f_4(\text{Grup}) + f_m(x_m) \qquad (7,8)$$

Acest model specifică o distribuție (cum ar fi o distribuție normală sau o distribuție binomială) și o **funcție de legătură** g care leagă valoarea așteptată a distribuției la m variabile predictoare, precum și încercările de a potrivi funcțiile $f_i(x_i)$ pentru a satisface Ecuația (7.8) .

Funcțiile *fi (xi) pot fi ajustate folosind mijloace parametrice sau neparametrice , oferind astfel* potențialul pentru potriviri mai bune la date decât alte metode. Cu toate acestea, metoda este foarte generală.

Utilizarea modelării regresiei pentru controlul confuziei

În această abordare a controlului confuziei, obiectivul modelării de regresie este de a descoperi efectul expunerii asupra bolii, în timp ce controlează sau ajustează pentru celelalte. Pașii procesului sunt următorii:

1. Începeți cu o ecuație model în care partea dreaptă include termenii care reprezintă expunerea și potențialii factori de confuzie.
2. Apoi, introduceți variabila cu cea mai semnificativă relație biostatistică cu rezultatul de sănătate în model.
3. În continuare, o a doua variabilă este aleasă din restul pe baza faptului că adaugă cea mai mare „semnificație" modelului creat în pasul anterior.
4. Continuați acest proces până când variabilele suplimentare nu vor crește semnificația în model.

Este în totalitate posibil, atunci când se utilizează pachete biostatistice disponibile, ca proceduri automate diferite să ducă la selecții diferite pentru același set de date. Întrucât niciun algoritm biostatistic singur nu poate identifica toți factorii de confuzie, va fi necesară o judecată epidemiologică bazată pe cunoștințele subiectului pentru a finaliza procesul decizional. Cu alte cuvinte, procedurile de selecție automatizate nu ar trebui utilizate pentru a identifica factorii de confuzie, deoarece aceste proceduri pot duce la includeri sau excluderi neadecvate într-un model.

MODIFICĂRI SUPLIMENTARE PENTRU CONTROLUL CONFUNDEI PRIN MODELARE DE REGRESIUNE

Au fost dezvoltate și practicate o serie de modificări pentru a controla în continuare confuzia prin modelarea regresiei. Urmează câteva exemple.

PRIN COMPARAREA GRUPURILOR. Când cineva este capabil să clasifice doar subiecții cazului în grupuri, cum ar fi *scăzut , mediu* și *ridicat ,* expunerea este **categorică** . Astfel de variabile pot fi clasificate ca predictori în partea dreaptă a ecuației model. De exemplu, luați în considerare un caz cu cele două categorii Expus și Neexpus, pe baza exemplului precedent pentru VEMS în două grupuri, în care intenționați să ajustați comparația pentru vârstă, înălțime și fumat. Setul de date conține o variabilă „Grup", care ia valoarea 1 dacă subiectul cazului este expus și 0 dacă subiectul cazului nu este expus. Modelul 2 al ecuației (7.5) poate fi apoi modificat astfel:

Model 6: VEMS = a + b (Vârsta) + c (Înălțime) + d (Țigări) + e (Grup) + r (7,9)

în care coeficientul e este de interes primordial. Reprezintă efectul creșterii variabilei Grup cu o unitate - presupunând că modelul este corect și că toate celelalte variabile sunt menținute constante. Este diferența medie în VEMS între a fi expus

(Grup = 1) și neexpuse (Grup = 0) atunci când Vârsta, Înălțimea și Țigaretele au fost

aceleași pentru fiecare grup.

MODELE DE REGRESIUNE LOGISTICĂ PENTRU REZULTATELE BOLII. În aceeași investigație epidemiologică, dacă măsura rezultatului este absența sau prezența unei boli respiratorii, se solicită o variabilă dihotomică cu valorile corespunzătoare 0 (pentru absența bolii) și 1 (pentru prezența bolii).

Probabilitatea p de îmbolnăvire este apoi o funcție de vârstă, pachetul cumulativ de ani de fumat și concentrația de expunere, pentru care un model poate fi:

$$\text{Model 7: } \ln\{p/(1-p)\} = a + b\,(\text{Vârsta}) + c\,(\text{Pack-Ani}) + d\,(\text{Concentrație}) \qquad (7.10)$$

Acesta este un **model de regresie logistică** . Termenul *logistic* reprezintă forma părții stângi a ecuației (7.10) . Se presupune că componenta aleatoare are o distribuție binomială.

MODELE DE REGRESIUNE DE OTRAVĂ PENTRU REZULTATELE BOLII. Pentru investigațiile de incidență în care durata urmăririi variază între subiecții cazului, un model de regresie logistică nu este adecvat. În astfel de investigații, parametrul de bază de interes este

$$\text{Rata de incidență} = \frac{\text{Numărul de cazuri de boală } Y}{\text{Persoană - Timp de observație } T}$$

Modelul de regresie Poisson este foarte potrivit pentru studierea acestor rate.

Luați în considerare un studiu de cohortă ocupațional cu două grupuri de expunere supuse controlului pentru cancerul pulmonar de-a lungul mai multor ani. Un model de regresie Poisson adecvat, cu predictori Vârsta, Anii de pachet de țigări și Grupul de expunere, poate lua următoarea formă:

$$\text{Model 8: } \ln(T) = a + b\,(\text{Vârsta}) + c\,(\text{Pack-Ani}) + d\,(\text{Grup}) + r \qquad (7.11)$$

Deoarece se presupune că componenta aleatorie a modelului urmează o distribuție Poisson , acest model se numește **model de regresie Poisson** .

Confuzie și coliniaritate

CE ESTE COLLINEARITATEA? (SPITALUL CHILDREN'S MERCY, 2000)

În biostatistică, **coliniaritatea** apare atunci când există o relație liniară *aproape* perfectă între unele sau toate variabilele independente dintr-un model de regresie. (Este cunoscută și sub denumirea de *proastă condiționare* , *aproape coliniaritate* sau *multicoliniaritate* .) Această situație poate duce la o anumită redundanță sau suprapunere între aceste variabile.

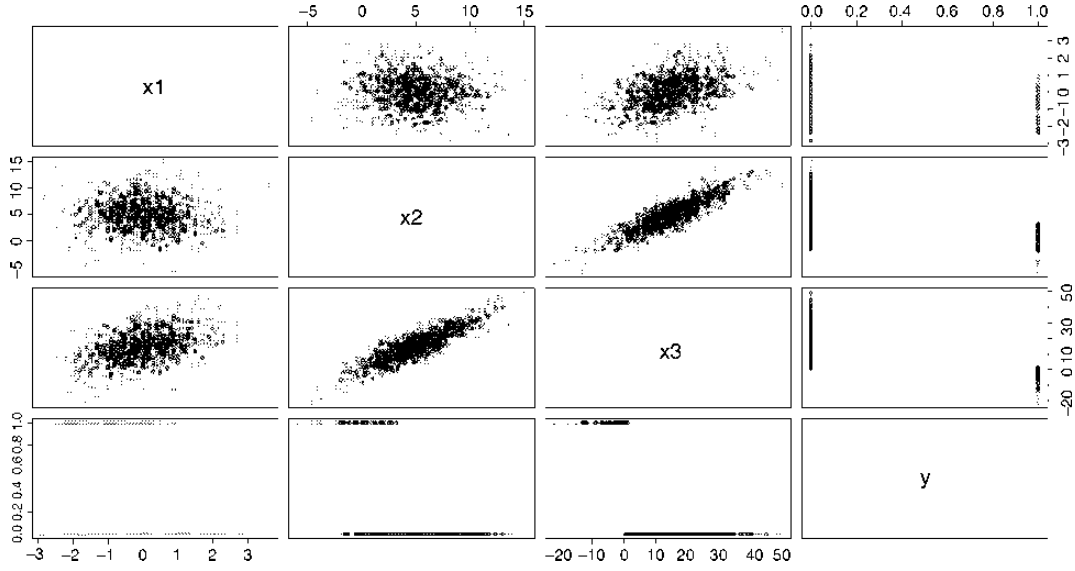
Coliniaritatea poate apărea ca o corelație foarte mare între două variabile independente. Un alt tip de coliniaritate are ca mai multe variabile să se adaugă la ceva care este foarte aproape de o valoare constantă.

Coliniaritatea poate provoca o pierdere a puterii biostatistice, făcând interpretarea mai dificilă. Următorul exemplu arată confuzie din cauza colinirii simple între două variabile aparent independente.

> **Exemplul 7.4:** Un exemplu de coliniaritate y[29]

Acest exemplu arată că pentru variabilele continue, poate apărea confuzie din cauza faptului că două variabile sunt coliniare cu o a treia, așa cum se arată în următorul segment de cod R:

```
> # Creați o primă variabilă dependentă x1 (rotunjire la o zecimală):
> x1 <- rotund(rnorm(1000, medie = 0, sd=1), 1)
> # Apoi, creați o a doua variabilă dependentă x2 (cu o zecimală exactă):
> x2 <- rotund(rnorm(1000, medie = 5, sd=3), 1)
> # În cele din urmă, creați o a treia variabilă independentă x3 care este o funcție directă a
> # x1 și x2:
> x3 <- 5*x1 + 3*x2
> # Acum, creați o variabilă de rezultat binară care depinde de toate cele trei variabile.
> # Rețineți că probabilitatea binomului este o funcție R inv.logit().
> y <- rbinom(1000, 1, exp(x1 + 3*x2 -5 * x3)/(1+ exp(x1 + 3*x2 -5 * x3)))
> collinear.dat <- data.frame(x1=x1, x2=x2, x3=x3, y=y)
```




```
> pairs(colinear.dat) # Ieșire: Figura 7.17 .

> # Analiză suplimentară pentru coliniaritate
```

FIGURE 7.17 Confounding analysis: high correlation is obvious, but assessing perfect

[29] www.medicine.mcgill.ca/epidemiology/Joseph/courses/EPIB-621/main.html

```
> ieşire <- glm(y ~ x1 + x2 + x3, data = colinear.dat,
+                    familie=binom)
```

Mesaj de avertizare:
glm.fit: au apărut probabilităţi ajustate numeric 0 sau 1

```
> # Aparent R a detectat coliniaritate; reţineţi mesajul de avertizare de mai jos:
> rezumat(ieşire) # Ieşire:
```

Apel: glm(formula = y ~ x1 + x2 + x3, family = binom, data = colinear.dat)

Reziduuri de devianta:

Min	1Q	Median	3Q	Max
-2.122	0.000	0.000	0.000	1.815

Coeficienţi: (1 nu este definit din cauza singularităţilor)

	Estima	Std. Eroare	valoarea z	Pr(>|z|)
(Intercepta)	0,8332	1,2312	0,677	0,49858
x1	-22,2919	6,9281	-3.218	0,00129
x2	-12,0320	3,7462	-3.212	0,00132
x3	N / A	N / A	N / A	N / A

Signif. coduri: 0 '*' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 " 1**

(Parametrul de dispersie pentru familia binomială considerat a fi 1)

Devianţă nulă: 517,564 pe 999 grade de libertate Devianţă reziduală: 18,205 pe 997 grade de libertate AIC: 24.205

Numărul de iteraţii Fisher Scoring: 14

```
> # x3 a fost acum eliminat; şi alte variabile sunt estimate în mod rezonabil.
> # Pentru a obţine automat CI-urile, rulaţi din nou modelul fără x3, folosind
> funcţia R încorporată confint():
> ieşire <- glm(y~x1+x2, data=colinear.dat.family = binom)
> # Ieşire:
```

Mesaj de avertizare:
glm.fit: au apărut probabilităţi ajustate numeric 0 sau 1

```
> # Acest mesaj indică rezultate puternice, dar nu coliniaritate.
> limită (ieşire)
```

Se aşteaptă ca profilarea să se facă...

	2,5 %	97,5 %
(Interceptar	-1,351273	3,736777
e) x1	-41,218618	-12,389357
x2	-22,278406	-6,683552

Au existat 36 de avertismente (utilizaţi avertismente() **pentru a le vedea)**

```
> # Verificarea câtorva valori ajustate:
> output$fitted[1:5]
```

```
12345
9.999559e-01    2.220446e-16  2.220446e-16    2.220446e-16    2.220446e-16
> # Notă: Unele valori sunt aproape de zero, în timp ce altele sunt mai mari.
```

Întrebări de revizuire pentru Secțiunea 7.3

1. (a) În investigațiile epidemiologice, ce este un factor de confuzie sau o variabilă de confuzie? (b) Dați două exemple de astfel de investigații și factorii de confuzie asociați.
2. Numiți două abordări biostatistice pentru controlul confuziei în investigațiile epidemiologice și descrieți pe scurt fiecare abordare, dând exemple.
3. Dacă confuzia nu este corect ajustată, care sunt posibilele consecințe pentru rezultatele analitice ale unui studiu epidemiologic? Dați câteva exemple.
4. Descrieți controalele combinate experimentale și analitice disponibile în mod obișnuit ale confuziei și dați exemple.
5. Descrieți metoda MH de stratificare și dați un exemplu despre cum este utilizată în controlul confuziei.
6. Descrieți utilizarea modelării de regresie în scopul ajustării confuziei într-o investigație biostatistică. Da un exemplu.
7. Scrieți cinci modele de regresie posibile potrivite pentru ajustarea confuziei în studiile de sănătate publică și dați exemple de utilizare a acestora.
8. (a) Ce este coliniaritatea? Da un exemplu.
   (b) Cum poate fi folosită funcția confint() pentru a testa coliniaritatea?
9. (a) În analiza de supraviețuire a datelor longitudinale, cum afectează confuzia calculul?
   (b) Cum poate fi ajustată confuzia într-o astfel de analiză?
10. (a) Când se utilizează modelul de regresie al hazardelor proporționale Cox pentru a se potrivi unui set de date, cum sunt ajustate datele pentru efectele de confuzie?
    (b) Dacă același set de date ar fi analizat folosind diagrama K–M, cum se controlează efectele de confuzie?

Exerciții pentru Secțiunea 7.3

1. *Analiza supraviețuirii: curbele caracteristice de funcționare a receptorului (ROC) în* R

.

Curba ROC (prezentată în Figura 7.18) este un instrument utilizat în predicție și clasificare . Acesta arată compromisul dintre sensibilitate și specificitate - pentru o serie de praguri aplicate unei variabile explicative - pentru a prezice o decizie binară a rezultatului . Relația dintre sensibilitate și specificitate este de obicei reprezentată grafic, aria de sub curbă fiind adesea folosită ca măsură a performanței discriminatorii: 1 = perfect și 0,5 = nu este mai bun decât o estimare aleatorie.

Funcția R survivalROC () , în pachetul CRAN survivalROC , creează curbe ROC dependente de timp din datele de supraviețuire cenzurate, folosind metoda K-M sau estimarea celui mai apropiat vecin (NNE) a lui Heagerty, Lumley și Pepe (2000) .

Formula de utilizare pentru funcția survivalROC() este survivalROC(Stime, status, marker, entry = NULL, predict.time, cut.values = NULL, method = "NNE", lambda = NULL, span = NULL, window = "simetric ") pentru care argumentele sunt

Stime	Ora evenimentului sau timpul de cenzurare pentru subiecte.
status	Indicator de stare; 1 dacă deces sau eveniment, 0 în caz contrar.

marker	Valoarea predictorului sau a marcatorului.
intrare	Ora de intrare pentru subiecte.
predict.reduc	Punctul temporal al curbei ROC.
ere de timp.valori	Valorile markerului de utilizat ca limită pentru calcularea sensibilității și specificității.
metodă	Metoda de montare a distribuției de îmbinare a (x, t), fie a lui K–M, fie a NNE; metoda implicită este NNE.
lambda	Parametru de netezire pentru NNE.
span	Interval pentru NNE; nevoie fie de lambda, fie de span
fereastră	pentru NNE. Fereastră pentru NNE, fie simetrică, fie

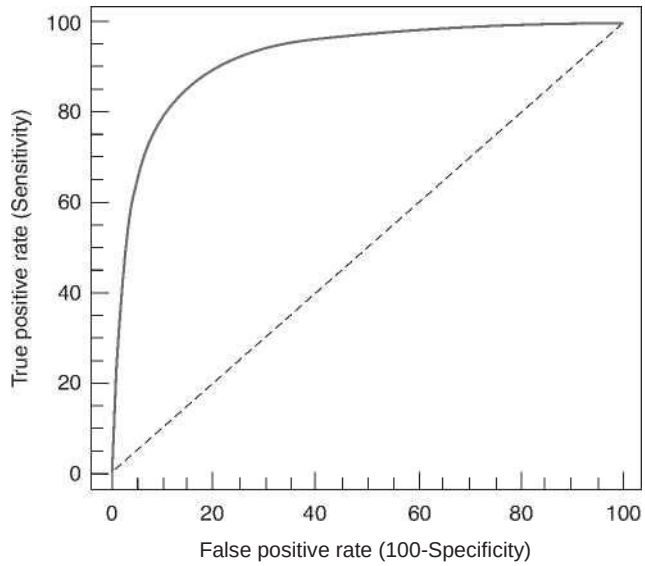


FIGURE 7.18 An ROC curve.

Să presupunem că ați cenzurat datele de supraviețuire împreună cu o valoare a markerului de referință și ați dori să vedeți cât de bine prezice markerul timpul de supraviețuire pentru subiecții din setul de date. În special, să presupunem că aveți timpi de supraviețuire în zile și doriți să vedeți cât de bine prezice markerul supraviețuirea pe 1 an (timpul estimat = 365 de zile). Funcția roc.KM.calc() returnează valorile unice ale markerului TP (adevărat pozitiv), FP (fals pozitiv), estimarea de supraviețuire K–M corespunzătoare punctului de timp de interes (predict.time) și AUC [area sub (curba ROC] la momentul de interes.

Funcția survivalROC() returnează o listă cu următoarele elemente:

cut.values	Valori unice ale markerului pentru calcularea TP și FP .
TP	Adevărat pozitiv corespunzător limitelor din *x* .

FP	Fals pozitiv corespunzător limitelor din *x* .
predict.time	Punct temporal de interes.
Survival AUC	Estimarea supraviețuirii K–M la predict.time .
	Aria de sub curba ROC la timp predict.time .

Folosind setul de date mayo , următorul segment de cod R este utilizat pentru a genera o curbă ROC ca în Figura 7.18 , folosind metoda NNE:

```
> install.packages ("survivalROC")
>  bibliotecă (survivalROC)
>  ls(„pachet:survivalROC")
```

[1] „survivalROC" „survivalROC.C"

```
>  date (mai)
```

maia

	timp	Cenzorul	mayoscore5	mayoscore4
1	**41**	**1**	**11.251850**	**10.629450**
2	**179**	**1**	**10,136070**	**10,185220**
3	**334**	**1**	**10.095740**	**9.422995**
312	**533**	**0**	**6,115321**	**6,902997**

```
>  nobs <- NROW (maia)
>  nobi
```

[1] 312

```
>  cutoff <- 365
>
>  ## MAYOSCORE 4, METODĂ = NNE:
>  Mayo4.1= survivalROC(Stime=mayo$time,
+ status=mayo$cenzor,
+ marker = mayo$mayoscore4,
+ predict.time = cutoff,span = 0,25*nobs^(-0,20) )
>  plot(Mayo4.1$FP, Mayo4.1$TP, type="l", xlim=c(0,1),
+ ylim=c(0,1),
+ xlab=paste("FP", "\n", "AUC = ",round(Mayo4.1$AUC,3)),
+ ylab="TP",main="Mayoscore 4, Method = NNE \n Year = 1")
>  abili (0,1)
>  # Ieșire: Figura 7.19 .
```

Mayoscore 4, Metodă = NNE
Anul = 1

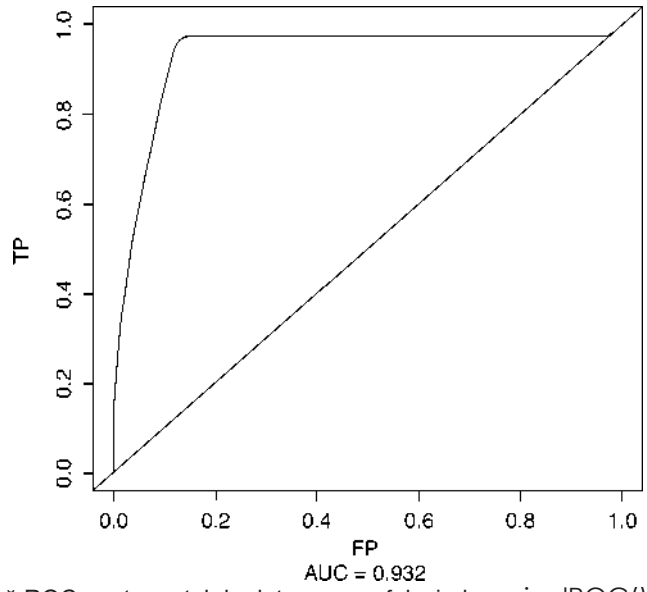


FIGURA 7.19 O curbă ROC pentru setul de date mayo folosind survivalROC() ; metoda = NNE (estimarea celui mai apropiat vecin).

(a) Explicați funcția fiecărei linii a segmentului de cod R pentru acest calcul.
(b) Reluați acest segment de cod în mediul R.
(c) Comentați complotul rezultat.

2. *Analiza supraviețuirii: curbele ROC în* R . Din nou, folosind setul de date mayo , următorul segment de cod R este utilizat pentru a genera o curbă ROC (Figura 7.20), dar folosind metoda K–M:

```
>
> ## MAYOSCORE 4, METODĂ = KM:
> Mayo4.2= survivalROC(Stime=mayo$time,
+ status=mayo$cenzor,
+ marker = mayo$mayoscore4,
+ predict.time = cutoff, method="KM")
> plot(Mayo4.2$FP, Mayo4.2$TP, tip="l", xlim=c(0,1),
+ ylim=c(0,1),
+ xlab=paste("FP", "\n", "AUC = ,round(Mayo4.2$AUC,3)),
+ ylab="TP",main="Mayoscore 4, Method = KM \n Year = 1")
>   abili (0,1)
>   # Ieșire: Figura 7.20 .
```

(a) Explicați funcția fiecărei linii a segmentului de cod R pentru acest calcul.
(b) Reluați acest segment de cod în mediul R.
(c) Comentați complotul rezultat.

Mayoscore 4, Metoda = KM
Anul = 1

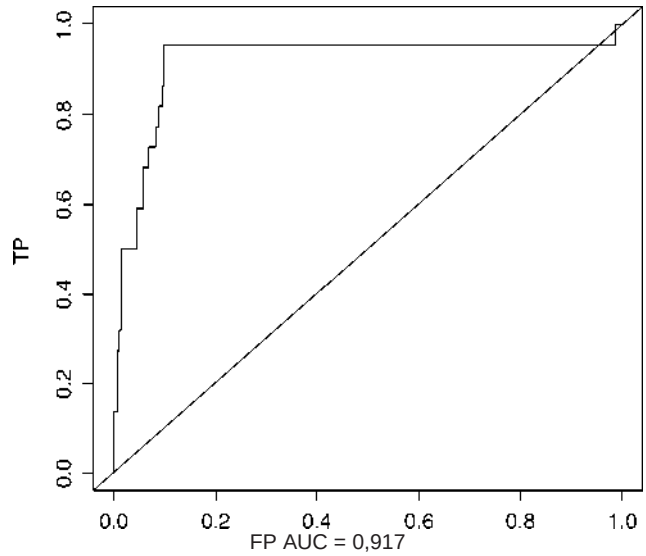


FIGURA 7.20 Curba ROC pentru setul de date mayo folosind survivalROC(); metoda = Kaplan–Meier (K–M).

3. *Analiza de supraviețuire și confuzie: pachetul CRAN* NestedCohort *pentru analiza de supraviețuire a investigațiilor epidemiologice imbricate în cohorte* (Katki & Mark, 2008, 2009) .

Acest pachet conține funcții R care efectuează analize de supraviețuire (modele K–M și Cox) pentru investigații epidemiologice imbricate în cohorte; aceste funcții sunt utilizate pentru estimarea ratelor hazardului, estimarea probabilităților de supraviețuire și standardizarea factorilor de confuzie. Funcția NestedCohort() este adecvată în special pentru evaluarea probabilităților de supraviețuire și a riscurilor cunoscute.

În următorul exemplu lucrat, observațiile au constat în rezultatul cancerului esofagian și timpul de supraviețuire pe toți subiecții cazului, împreună cu factori de confuzie cunoscuți. Principalul interes investigațional sa concentrat pe efectul concentrațiilor diferitelor metale, în special zincul, asupra acestui cancer. Cu toate acestea, măsurarea efectului concentrațiilor a necesitat țesut de biopsie esofagiană și o tehnică de măsurare costisitoare și a fost dificil și costisitor măsurarea concentrațiilor pe toți subiecții cazului. Prin urmare, anchetatorii au măsurat concentrația de zinc (împreună cu cupru, nichel, fier, calciu și sulf) pe o probă selectată din cohortă. Acest eșantion a supraeșantionat cazurile și pe cei cu histologii inițiale avansate (adică, cei mai probabil să devină cazuri), deoarece aceștia au fost considerați cei mai informați subiecti de caz. Din cauza constrângerilor de disponibilitate și de cost, mai puțin de 30% din cohortă a putut fi eșantionată.

În acest exemplu, funcția NestedCohort() furnizează rate de pericol ajustate, probabilități de supraviețuire standardizate și riscuri atribuibile populației (PAR) pentru efectul zincului asupra cancerului. În special, în acest exemplu, funcția nested.km() este utilizată pentru a estima și potrivi curbele de supraviețuire K–M la datele cohortei imbricate.

Segmentul de cod R pentru analiză este următorul:

```
> install.packages(„NestedCohort")
> bibliotecă (NestedCohort)
```

Se încarcă pachetul necesar: supraviețuire
Se încarcă pachetul necesar: spline
Se încarcă pachetul necesar: MASS

```
> ls(„pachet:NestedCohort")
```

[1] „nested.coxph" „nested.km" „nested.stdsurv"

```
> date (zinc)
> atașați (zinc)
> zinc # Examinând detaliile setului de date.:
```

	id8	Sex	varsta	agestrfum	bea	anyhist
1	10100012	Femeie	53	51<=Vârsta<=6	NuNu	<NA>
2	10100123	Femeie	54	51<=Vârsta<=6	NuNu	<NA>
3	10300066	Bărb	54	51<=Vârsta<=6	Nu Vreodat	<NA>

440 32500344 Femeie 57 51<=Vârsta<=60 Niciodată Niciodată Nicio familie
Istorie

	bazeist	disp1	disp2	mildysp	moddysp
1	Normal	1	0	Cel mai rău nu	Cel mai rău nu este
2	Normal	1	0	Cel mai rău nu	Cel mai rău nu este
3	Normal	1	0	Cel mai rău nu	Cel mai rău nu este

440 Normal 1 0 Cel mai rău nu este ușor Cel mai rău nu este moderat

	Sevdysp	ec01	futime01	zincset	la sută
1	Cel mai rău nu	0	5980	Elemente neobservate	N / A
2	Cel mai rău nu	0	5980	Elemente neobservate	N / A
3	Cel mai rău nu este grav	0	5980	Elemente neobservate	N / A
440	Cel mai rău nu este grav	1	3973	Elemente observate	-0,02219326

	Aromă	cacent	gard	simpatic	cucent
1	N / A	N / A	N / A	N / A	N / A
2	N / A	N / A	N / A	N / A	N / A
3	N / A	N / A	N / A	N / A	N / A
440	0,187727832	-1,08682017	0,662201065	-0,787517392	0,811040423

	zncent	pqt	sqt	caqt	feqt	niqt	cuqt	znqt	pq1	pq2	pq3	pq4	sq1	sq2	sq3
1	NA	NANA		NANANA		NANA		NA	NANA		NANANA		NANA		
2	NA			NA	NANA	NA	NANA			NANANA		NANA			
3	NA			NA	NANA	NA	NANA			NANANA		NANA			
440	0.680324598	2		3		1	42330			100001					

	sq4	caq1	caq2	caq3	caq4	feq1	feq2	feq3	feq4	niq1	niq2	niq3	niq4	cuq1	cuq2
1	NANA	N / A	N / A	N / A	N / A	NANA	N / A	NANA	NANA	NANA	N / A	N / A	N / A	N / A	N / A
2	NANA	N / A	N / A	N / A	N / A	NANA	N / A	NANA	NANA	NANA	N / A	N / A	N / A	N / A	N / A
3	NANA	N / A	N / A	N / A	N / A	NANA	N / A	NANA	NANA	NANA	N / A	N / A	N / A	N / A	N / A
440	01		0	0	0	00		0	101		00		0	0	

	cuq3	cuq4	znq1	znq2	znq3	znq4	stdagepill	znquartiles	observat
1	N / A	N / A	N / A	N / A	N / A	N / A	-0,1818182	<NA>	0
2	N / A	N / A	N / A	N / A	N / A	N / A	0,0000000	<NA>	0
3	N / A	N / A	N / A	N / A	N / A	N / A	0,0000000	<NA>	0
440	1	0	0	0	1	0	0,5454545	Q3	1

> str(zinc)

'data.frame' : 431 obs. din 61 de variabile:

$ id8 : int 10100012 10100123 10300066 10400038 10400106 10400245 10500252 10500267 10800011 10800049 ...

$ sex : Factor cu 2 niveluri „Femeie”, „Bărbat”: 1 1 2 2 2 1 1 1 2 2 ...

$ agepill : int 53 54 54 44 44 43 49 48 41 61 ...

$ agestr : Factor cu 3 niveluri „Vârsta<=50”,”,51<=Vârsta<=60”, ..: 2 2 2 1 1 1 1 1 1 3

$ smoke Factor cu 2 niveluri „Niciodată”, „Niciodată”: 1 1 1 1 1 1 1 1 2 ...

$ drink Factor cu 2 niveluri „Niciodată”, „Niciodată”: 1 1 2 2 1 1 1 1 2 2 ...

$ anyhist Factor cu 2 nivele „Fără istoric familial”,..: NA NA NA NA NA NA NA NA NA ...

```
$ basehist   Factor cu 7 niveluri „Normal”, „Esofagita”, ..: 1 1 1 1 3 2 1 1 1 1 ...
$ dysp1      int 1 1 1 1 3 2 1 1 1 1 ...
$ dysp2        int 0 0 0 0 1 0 0 0 0 0 ...
$ mildysp      Factor cu 2 niveluri „Cel mai rău nu este ușor”, ..: 1 1 1 1 2 1 1 1 1 ...
$ moddysp    Factorul cu 2 niveluri „Cel mai rău nu este moderat”, ..: 1 1 1 1 1 1 1 1 1 ...
$ sevdysp    Factorul cu 2 niveluri „Cel mai rău nu este sever”, ..: 1 1 1 1 1 1 1 1 1 ... num 0 0 0 0 0 0 0 0 0 0 ...
$ ec01       int 5980 5980 5980 5980 5980 3404 5980 5980 5980 5980 ...
$ futime01   Factor cu 2 niveluri „Elemente neobservate”, ..: 1 1 1 1 1 1 1 1 1 1 ...
$ zincset
$ pcent      num NA NA NA NA NA NA NA NA NA NA ...
$ scent      num NA NA NA NA NA NA NA NA NA NA ...
$ cacent     num NA NA NA NA NA NA NA NA NA NA ...
$ fecent     num NA NA NA NA NA NA NA NA NA NA ...
$ nicent     num NA NA NA NA NA NA NA NA NA NA ...
$ cucent     num NA NA NA NA NA NA NA NA NA NA ...
$ zncent     num NA NA NA NA NA NA NA NA NA NA ...
$ pqt          int NA NA NA NA NA NA NA NA NA NA ...
$ sqt          int NA NA NA NA NA NA NA NA NA NA ...
$ caqt         int NA NA NA NA NA NA NA NA NA NA ...
$ feqt         int NA NA NA NA NA NA NA NA NA NA ...
$ niqt         int NA NA NA NA NA NA NA NA NA NA ...
$ cuqt         int NA NA NA NA NA NA NA NA NA NA ...
$ znqt         int NA NA NA NA NA NA NA NA NA NA ...
$ pq1          int NA NA NA NA NA NA NA NA NA NA ...
$ pq2          int NA NA NA NA NA NA NA NA NA NA ...
$ pq3          int NA NA NA NA NA NA NA NA NA NA ...
$ pq4          int NA NA NA NA NA NA NA NA NA NA ...
$ sq1          int NA NA NA NA NA NA NA NA NA NA ...
$ sq2          int NA NA NA NA NA NA NA NA NA NA ...
$ sq3          int NA NA NA NA NA NA NA NA NA NA ...
$ sq4          int NA NA NA NA NA NA NA NA NA NA ...
$ caq1         int NA NA NA NA NA NA NA NA NA NA ...
$ caq2         int NA NA NA NA NA NA NA NA NA NA ...
$ caq3         int NA NA NA NA NA NA NA NA NA NA ...
$ caq4         int NA NA NA NA NA NA NA NA NA NA ...
$ feq1         int NA NA NA NA NA NA NA NA NA NA ...
$ feq2         int NA NA NA NA NA NA NA NA NA NA ...
$ feq3         int NA NA NA NA NA NA NA NA NA NA ...
$ feq4         int NA NA NA NA NA NA NA NA NA NA ...
$ niq1         int NA NA NA NA NA NA NA NA NA NA ...
$ niq2         int NA NA NA NA NA NA NA NA NA NA ...
$ niq3         int NA NA NA NA NA NA NA NA NA NA ...
$ niq4         int NA NA NA NA NA NA NA NA NA NA ...
$ cuq1         int NA NA NA NA NA NA NA NA NA NA ...
$ cuq2         int NA NA NA NA NA NA NA NA NA NA ...
```

```
$ cuq3          : int NA NA NA NA NA      NA NA       NA    NA NA ...
$ cuq4          : int NA NA NA NA NA      NA NA       NA    NA NA ...
$ znq1          : int NA NA NA NA NA      NA NA       NA    NA NA ...
$ znq2          : int NA NA NA NA NA      NA NA       NA    NA NA ...
$ znq3          : int NA NA NA NA NA      NA NA       NA    NA NA ...
$ znq4          : int NA NA NA NA NA      NA NA       NA    NA NA ...
$ sttagepill :  num -0,182 0 0 -1,818 -1,818 ...
$znquartiles :  Factor cu 4 niveluri "Q1","Q2","Q3",..: NA NA NA NA NA NA NA
                NA NA NA ...
$ observat:     num 0 0 0 0 0 0 0 0 0 0 ...

>
>   mod <- imbricat.km(survfitformula =
+                          "Surv(futime01,ec01==1)~znquartiles",
+ samplingmod = "ec01*basehist", expuneredeinteres = "Q4",
+ date = zinc)
```

Diferențele de risc față de znquartiles=Q4 după ora 5980

	Diferența de risc	StdErr	95% CI Stânga	95% CI
Q4 - Q1	0,2817534	0,10416236	0,07759516	0,4859116
Q4 - Q2	0,0555103	0,07565667	-0,09277677	0,2037974
Q4 - Q3	0,1068147	0,08073547	-0,05142680	0,2650562

```
>   rezumat(mod)
```

Apel: survfit(formula = as.formula(survfitformula), date = date, ponderi = 1/pihat, na.action = na.omit, type = "fl")

308 observații șterse din cauza lipsei znquartiles=Q1

Timp	n.risc	n.eveni	supraviețț	std.err	IC mai mic de	95% CI
1037	59,8	1.42	0,977	0,0235	0,840	0,997
4143	44.4	1.42	0,946	0,0388	0,789	0,987
5189	41.1	1,37	0,915	0,0514	0,736	0,975
5893	59,8	1,57	0,633	0,0862	0,441	0,775

znquartiles=Q2

Timp	n.risc	n.eveni	supraviețț	std.err	IC mai mic de	95% CI
1038	116,9	1,57	0,987	0,0133	0,909	0,998
1064	115.3	4,51	0,949	0,0260	0,864	0,981
1070	110,8	2.33	0,929	0,0324	0,830	0,971
4139	63,5	1,37	0,859	0,0520	0,718	0,933

znquartiles=Q3

timp	n.risc	n.eveni	supravie	std.err	IC mai mic de	95% CI
318	125.1	1.20	0,990	0,00948	0,934	0,999
733	123,9	1.20	0,981	0,01340	0,926	0,995
1001	122,7	1,37	0,970	0,01759	0,907	0,991
5351	64,6	1,42	0,808	0,05800	0,662	0,896

znquartiles=Q4

Timp	n.risc	n.eveni	supravie	std.err	IC mai mic de	95% CI
1037	59,8	1.42	0,977	0,0235	0,840	0,997
4143	44.4	1.42	0,946	0,0388	0,789	0,987
5189	41.1	1,37	0,915	0,0514	0,736	0,975

```
> plot(mod, ymin = 0,6, xlab = "Timp în zile",
+              ylab = "Probabilități de supraviețuire",
+ principal = "Analiză de supraviețuire prin Quartila de zinc",
+              lty = 1:4)
> legenda(1000,0.7, c("Q1", "Q2", "Q3", "Q4"), lty=1:4)
> # Ieșire: Figura 7.21 .
```

(a) Explicați funcția fiecărei linii a segmentului de cod R pentru acest calcul.
(b) Reluați acest segment de cod în mediul R.
(c) Comentați graficul rezultat (Figura 7.21).
(d) În această analiză, cum este ajustată confuzia?

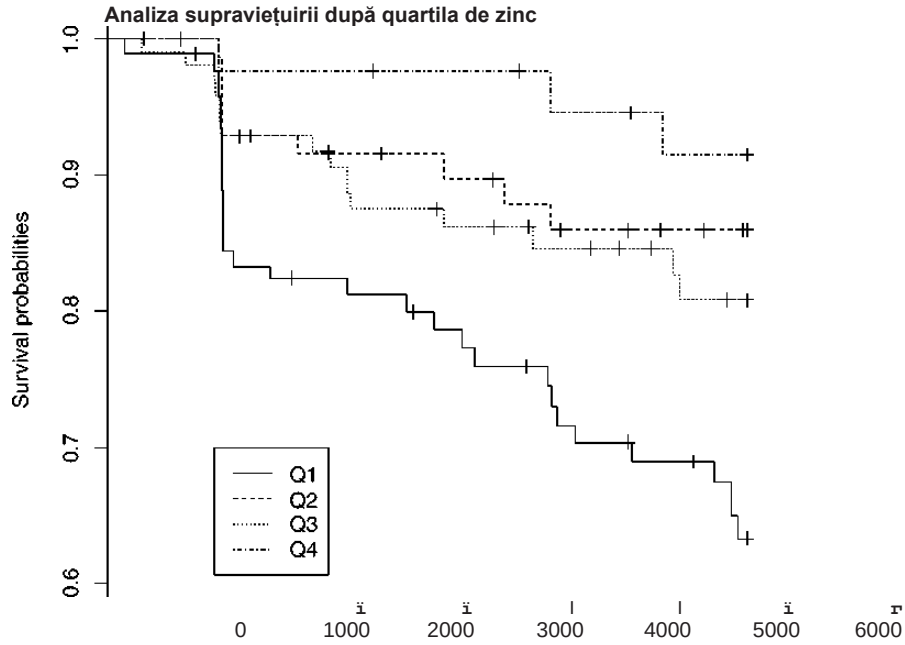


FIGURA 7.21 Curbele de analiză a supraviețuirii K–M pentru supraviețuirea cancerului pentru fiecare cuartilă de zinc, standardizate pentru factori de confuzie.

4. *Analiza supraviețuirii: pachetul CRAN* survivalBIV *(funcția de distribuție bivariată)* (Moreira et al, 2012) .

Pachetul CRAN survivalBIV estimează funcția de distribuție bivariată pentru evenimente ordonate secvenţial sub cenzură univariată. Acest pachet conține o serie de funcții R concepute pentru aplicații speciale în analiza de supraviețuire. Unele dintre aceste aplicații sunt ilustrate aici.

(a) Un estimator condiționat Kaplan–Meier (CK–M) : bivCKM() . Această funcție oferă estimări pentru funcția de distribuție bivariată pentru estimatorul CK–M și are o formulă de utilizare de forma:

```
bivCKM(obiect, t1, t2, conf = FALSE, n.boot = 1000,
conf.level = 0,95, method.boot = „percentilă")
```

cu urmatoarele argumente:

obiectul t1	Un obiect din clasa survBIV.
	Prima dată pentru obținerea estimărilor pentru funcția de distribuție bivariată; dacă lipsește, se va folosi 0 .
t2	A doua oară pentru obținerea de estimări pentru funcția de distribuție bivariată ; dacă lipsește, se va folosi timpul maxim2 .
Conf n.boot	Oferă benzi de încredere punctuale; implicit la FALSE .
conf.level	Numărul de mostre de bootstrap; implicit la 1.000 de mostre.
method.boot	Nivel de încredere; implicit la 0,95 (corespunzător la 95%).
	Metodă utilizată pentru a calcula CI bootstrap. Opțiunile posibile sunt „per centile" și „de bază"; implicit la „percentilă".

Următorul segment de cod R ilustrează utilizarea acestei funcții atunci când este aplicată setului de date bladderBIV (conținut în pachetul survivalBIV):

```
> install.packages("survivalBIV")
```

```
> library(SurvivalBIV)
```
Se încarcă pachetul necesar: prodlim
Se încarcă pachetul necesar: KernSmooth
KernSmooth 2.23 a fost încărcat
Copyright MP Wand 1997-2009

```
> ls("pachet:survivalBIV")
```

[1]	**"bivCKM"**	**"bivIPCW"**	**"bivKMPW"**
[4]	**"bivKMW"**	**"vezica**	**"corrBIV"**
[7]	**"dgpBIV"**	**„is.survBIV"**	**"plot.survBIV"**
[10]	**"rezumat"**	**"survBIV"**	**"survBIV"**

```
> date("bladderBIV")
> bladderBIV_obj <- cu(bladderBIV, survBIV(time1, eventi,
+ timp2, eveniment2))
> bivCKM(obiect = bladderBIV_obj, ti = 5, t2 = 20)
```
0,1920702
```
> #sau
> bladderBIV_obj <- survBIV(bladderBIV$timei,
+                           bladderBIV$eventi, bladderBIV$time2,
+                           bladderBIV$event2)
> bivCKM(obiect = bladderBIV_obj, ti = 5, t2 = 20, conf = TRUE,
+ nivel conf. = 0,9)
```

	5%	**95%**
0,1920702	**0,1185156**	**0,2657545**

(a) Explicați funcția fiecărei linii a segmentului de cod R pentru acest calcul.
(b) Reluați acest segment de cod în mediul R.
(c) Comentează rezultatele.
(d) O altă funcție R din acest pachet este bivIPCW() , care oferă estimări pentru funcția de distribuție bivariată pentru estimatorul de probabilitate inversă de cenzurare ed (IPCW). Următorul segment de cod R aplică această funcție la același set de date bladderBIV :

```
> date("bladderBIV")
> bladderBIV_obj <- cu(bladderBIV, survBIV(timei, eventi,
>                                  timp2, eveniment2))
> bivIPCW(obiect = bladderBIV_obj, ti = 5, t2 = 20,
>              method.cens = „prodlim")
```

Rulați acest segment de cod R și comentați rezultatul: **0,1868247**

```
> #sau
> bladderBIV_obj <- survBIV(bladderBIV$time1,
+ bladderBIV$event1, bladderBIV$time2,
+ bladderBIV$event2)
> bivIPCW(obiect = bladderBIV_obj, t1 = 5, t2 = 20,
+              conf = TRUE, conf.level = 0,9,
+ method.boot = „de bază")
```

```
                    5%          95%
0,1866335     0,1105516   0,2646384

>  date(„vezica urinarăBIV")
>  bladderBIV_obj <- cu(bladderBIV, survBIV(time1, event1,
+                              timp2, eveniment2))
>  bivKMPW(obiect = bladderBIV_obj, t1 = 5, t2 = 20) 0,1897386
>  # sau
>  bladderBIV_obj <- survBIV(bladderBIV$time1,
+                        bladderBIV$event1, bladderBIV$time2,
+                        bladderBIV$event2)
>  bivKMPW(obiect = bladderBIV_obj, t1 = 5, t2 = 20,
+            conf = TRUE, conf.level = 0,9)

                  5%          95%
0,1897386     0,1202276   0,2619701

>  date(„vezica urinarăBIV")
>  bladderBIV_obj <- cu(bladderBIV, survBIV(time1, event1,
+                              timp2, eveniment2))
>  bivKMW(obiect = bladderBIV_obj, t1 = 5, t2 = 20) 0,1921058
>  # sau
>  bladderBIV_obj <- survBIV(bladderBIV$time1,
+                        bladderBIV$event1, bladderBIV$time2,
+                        bladderBIV$event2)
>  bivKMW(obiect = bladderBIV_obj, t1 = 5, t2 = 20,
>            conf = TRUE, conf.level = 0,9)

                   5%          95%
0,1921058     0,1180276   0,2656078

>  # Exemplu pentru distribuția exponențială bivariată:
>  dgpBIV(n = 100, corr = 1, dist = „exponențial",
+ model.cens = „uniformă",
>  cens.par = 3, dist.par = c(1, 1), to.data.frame = TRUE)
```

	timp1	evenim	timp2	eveni
1	0,0005895626	1	0,75564858	1
2	0,6982030278	1	0,23072713	1
3	0,0558737679	1	0,85410701	1
4	0,6485372734	1	0,52543358	0
5	0,6713170423	0	0,00000000	0
6	0,8790889907	1	0,03043959	1
7	0,6068865498	1	0,10316399	1
8	0,0975181995	1	0,95176635	1
9	0,7997097191	0	0,00000000	0
10	0,2697271148	0	0,00000000	0
11	0,4367546837	0	0,00000000	0
12	0,2809143670	1	1,17781751	0
13	1.2135338530	1	0,10468994	1
14	1.4306760170	1	0,56626825	0
15	0,1780562948	0	0,00000000	0
16	0,3228512866	1	0,99409678	0
17	0,3177933234	1	0,99625677	0
18	0,2284733959	1	0,23741045	0
19	0,5274914715	1	0,30927187	1
20	0,2926479652	1	0,87673844	1
21	0,5181994882	1	0,84130949	1
22	1.6557341524	1	0,84904600	0
23	0,2561566898	1	0,02706028	1
24	1.0438891871	1	0,69471175	0
25	0,1166925130	0	0,00000000	0
26	0,2389276870	1	0,93371568	1
27	0,9899928542	1	1,72177471	0
28	0,2674041729	1	0,04035089	1
29	0,0177538705	0	0,00000000	0
30	0,7928811121	1	0,11645234	1
31	0,4532970815	1	0,81096417	1
32	0,9278468457	0	0,00000000	0
33	0,3206757427	1	0,27408256	1
34	0,2955963112	1	0,25176147	0
35	0,4530449047	1	0,12346276	1
36	1.9161922988	0	0,00000000	0
37	0,8300401610	0	0,00000000	0
38	0,2012449917	1	0,25938102	1

39	0,3447902738	1	0,02007339	1
40	0,9947645486	1	0,45997132	0
41	0,2069979079	1	0,28359194	1
42	0,5122642969	0	0,00000000	0
43	0,6378745642	1	0,06128698	1
44	1.9400842474	0	0,00000000	0
45	0,6603214571	0	0,00000000	0
46	0,6634516153	0	0,00000000	0
47	0,1856409729	1	0,23703461	1
48	0,9697090527	0	0,00000000	0
49	0,3939246334	1	0,12125044	0
50	0,0515332760	1	0,07251975	1
51	2.0156175639	0	0,00000000	0
52	1.2559016843	1	0,04350289	0
53	1.3967063204	1	0,18671369	1
54	0,9279630573	1	0,08916433	0
55	0,8164238634	1	0,05462287	1
56	0,0245129673	0	0,00000000	0
57	0,5157781137	1	0,26861503	1
58	0,3728056342	1	0,07303738	0
59	0,6157245501	1	0,41411110	0
60	0,4798975375	1	0,36525051	1
61	0,7601117366	0	0,00000000	0
62	0,4095801464	1	0,87768139	1
63	0,6044178131	1	0,04452789	0
64	0,0403524034	0	0,00000000	0
65	0,8983658056	0	0,00000000	0
66	0,4816842741	1	0,09519841	0
67	0,0310373344	1	0,36791676	1
68	0,9049734917	1	0,26968932	1
69	1.0606460375	1	1,37214910	1
70	0,1874751560	1	0,54789536	1
71	0,2055806991	1	0,94832239	1
72	1,8742257392	0	0,00000000	0
73	2.5159698048	1	0,26259136	1
74	2.6107399651	1	0,01456952	0
75	0,6080606624	0	0,00000000	0
76	2.2104980203	0	0,00000000	0
77	0,6895901640	0	0,00000000	0

78	0,1329671829	0	0,00000000	0
79	0,1822175515	1	0,13577832	1
80	0,6162270399	1	0,31167067	1
81	1.1962075245	0	0,00000000	0
82	1,0517517613	1	0,30542583	0
83	0,2711551202	1	0,93749705	0
84	0,4736238314	1	0,42495908	1
85	0,1668557518	0	0,00000000	0
86	0,4283004184	1	0,13572692	1
87	1.4244481432	1	0,25146951	1
88	0,2768717642	1	0,27988718	1
89	0,0087164905	1	0,03962044	1
90	2.4277168578	0	0,00000000	0
91	0,2097113177	0	0,00000000	0
92	0,0021386391	1	0,16532315	1
93	0,6734751035	1	1,46594482	0
94	1.8577060935	0	0,00000000	0
95	0,1409383491	0	0,00000000	0
96	0,7416686579	1	0,83971735	1
97	2.0469178744	0	0,00000000	0
98	0,5584133468	1	0,58440127	0
99	2.0667516228	1	0,07517950	1
100	0,2160539478	0	0,00000000	0

```
>   # Exemplu pentru distribuția bivariată Weibull:
>   dgpBIV(n = 100, corr = 1, dist = "weibull", model.cens =
+ „exponențial",
+ cens.par = 0,08, dist.par = c(2, 7, 2, 7))
```

$date

	timp1	evenime	timp2	evenim	Stime
1	1,388147729	0	0,0000000	0	1,388147729
2	0,743709478	0	0,0000000	0	0,743709478
3	4.311593760	1	2,2516124	0	6.563206116
4	4.470056068	0	0,0000000	0	4.470056068
5	5,997286522	0	0,0000000	0	5,997286522
6	0,005249357	0	0,0000000	0	0,005249357
7	2.609313119	0	0,0000000	0	2.609313119
8	7.146119245	1	0,3176864	0	7.463805680
9	6.447898338	1	5,9024270	1	12.350325305

10	4.169294109	1	4,2948521	1	8.464146250
11	8.247212600	1	4,7068117	0	12.954024341
12	3,595728674	1	6,0022950	1	9.598023643
13	10.117623649	1	3,7784823	0	13.896105950
14	8.209468634	0	0,0000000	0	8.209468634
15	7.006549462	1	9,0528632	1	16.059412663
16	6.835012301	1	2,7363197	0	9,571331956
17	8.472035429	1	3,5261148	0	11,998150183
18	2.849277534	1	8,8820817	1	11.731359253
19	3.051419189	1	6,6459485	1	9,697367718
20	6.553231439	1	0,7711965	1	7.324427896
21	11.004396078	1	6,2054960	1	17.209892032
22	2.567791299	1	7,7498997	0	10.317690996
23	4.204659119	1	10.1619853	1	14.366644423
24	7.360276219	0	0,0000000	0	7.360276219
25	2.734720899	0	0,0000000	0	2.734720899
26	2.904995325	1	6,7961131	1	9.701108398
27	11.153064089	1	1,6165403	1	12.769604432
28	6.035556386	1	1,7032070	1	7,738763414
29	7.222186275	1	7,4833385	1	14.705524741
30	5.468072242	0	0,0000000	0	5.468072242
31	12.965649468	1	3.1403897	1	16.106039130
32	3.061703269	1	0,7260095	0	3,787712725
33	3.751140080	1	9,1630566	0	12.914196669
34	1.680568964	1	3,8075400	0	5.488108939
35	4,353856486	1	6,9240779	1	11.277934414
36	0,572074048	0	0,0000000	0	0,572074048
37	6.673796084	1	5,2544671	1	11.928263191
38	7,689879433	1	7,5970093	1	15.286888712
39	10,583878851	1	2,0478638	0	12.631742695
40	4,089472140	0	0,0000000	0	4,089472140
41	11.093751137	1	0,9952871	0	12.089038279
42	0,756349065	0	0,0000000	0	0,756349065
43	2,838233864	0	0,0000000	0	2,838233864
44	5.285847816	0	0,0000000	0	5.285847816
45	5,899664003	1	5,8729714	1	11.772635387
46	5,924026349	1	1.2014802	1	7.125506594
47	2.294302556	1	5,7817984	1	8.076100916
48	7,228772283	1	0,8191344	0	8.047906647

49	6.849679163	1	4,2819505	1	11.131629712
50	6.960093893	1	6,7772215	0	13.737315346
51	3.703865355	1	13.3599371	1	17.063802422
52	2,978508038	0	0,0000000	0	2,978508038
53	0,632452127	0	0,0000000	0	0,632452127
54	4,437267513	0	0,0000000	0	4,437267513
55	11.433835033	0	0,0000000	0	11.433835033
56	0,026654385	0	0,0000000	0	0,026654385
57	4.366620830	1	4,6884216	1	9.055042468
58	2.086997694	0	0,0000000	0	2.086997694
59	4,386038504	1	4,1407420	0	8.526780544
60	0,166223885	0	0,0000000	0	0,166223885
61	5,881654616	1	4,8989473	1	10.780601869
62	4,387691111	0	0,0000000	0	4,387691111
63	3.327461537	0	0,0000000	0	3.327461537
64	5.439083596	1	5,9484436	1	11.387527240
65	3.043832240	0	0,0000000	0	3.043832240
66	3.866050823	0	0,0000000	0	3.866050823
67	9.206512038	1	1,9021350	1	11.108647010
68	2.077167493	0	0,0000000	0	2.077167493
69	1.169742962	0	0,0000000	0	1.169742962
70	7,836676569	0	0,0000000	0	7,836676569
71	3.094922049	1	7,5165716	1	10.611493699
72	5.456891474	1	3,0843889	0	8.541280357
73	6,777679345	1	1,6365333	1	8.414212682
74	4.037906132	1	6,6897972	1	10.727703306
75	0,513964477	0	0,0000000	0	0,513964477
76	6.690673682	1	1,0181006	1	7,708774302
77	13.279659553	1	13.3269520	1	26.606611569
78	3,841746730	1	4,1038628	1	7.945609505
79	1.444641809	0	0,0000000	0	1.444641809
80	2.937351998	1	2,8168339	0	5,754185881
81	1,223288129	1	6.1965012	1	7.419789348
82	1.275968027	0	0,0000000	0	1.275968027
83	3.611468174	0	0,0000000	0	3.611468174
84	6.549168425	1	1,3645790	1	7,913747386
85	7.801177148	1	3,5133309	1	11.314508024
86	6.137463114	1	16,4784673	1	22.615930434
87	7.846500803	0	0,0000000	0	7.846500803

88	**3.163988848**	**1**	**9,5916571**	**1**	**12.755645964**
89	**2.120092746**	**1**	**6,5040954**	**1**	**8.624188137**
90	**0,664639455**	**0**	**0,0000000**	**0**	**0,664639455**
91	**3,582123312**	**0**	**0,0000000**	**0**	**3,582123312**
92	**0,886321417**	**0**	**0,0000000**	**0**	**0,886321417**
93	**6.554446964**	**1**	**0,3444679**	**0**	**6,898914900**
94	**5,485892578**	**1**	**2,4127827**	**1**	**7,898675250**
95	**0,434836076**	**0**	**0,0000000**	**0**	**0,434836076**
96	**3.084640881**	**1**	**3,4676929**	**0**	**6.552333769**
97	**7.621512087**	**1**	**12,4836850**	**1**	**20.105197113**
98	**3.049349878**	**0**	**0,0000000**	**0**	**3.049349878**
99	**7.862905945**	**1**	**9,2293236**	**1**	**17.092229502**
100	**0,869275234**	**0**	**0,0000000**	**0**	**0,869275234**

```
attr(,"clasa")
[1] survBIV"
```

```
>
```

5. *Plot metode pentru un obiect survBIV.*

Următoarele segmente de cod R oferă diagramele pentru funcția de distribuție bivariată și distribuția marginală, cum ar fi cele din Exercițiul 4:

```
> install.packages("survivalBIV")
> bibliotecă (survivalBIV)
```

Se încarcă pachetul necesar: prodlim
Se încarcă pachetul necesar: KernSmooth KernSmooth 2.23 încărcat
Copyright MP Wand 1997-2009

```
> ls("pachet:survivalBIV")
```

[1]	**„bivC**	**"bivIPCW"**	**"bivKMPW"**
[4]	**"bivKMW"**	**"vezica**	**"corrBIV"**
[7]	**"dgpBIV"**	**„is.survBIV"**	**"plot.survBIV"**
[10]	**"rezumat"**	**"survBIV"**	**"survBIV"**

```
> date(„vezica urinarăBIV")
> bladderBIV_obj <- cu(bladderBIV, survBIV(time1, event1,
+                          timp2, eveniment2))
>
> op <- par(mfrow = c(2, 2))
> plot(bladderBIV_obj, plot.marginal = TRUE, method = "CKM")
```

Se așteaptă confirmarea modificării paginii...

```
> plot(bladderBIV_obj, plot.marginal = TRUE, method = "IPCW")
> plot(bladderBIV_obj, plot.marginal = TRUE, method = "KMPW")
```

```
> plot(bladderBIV_obj, plot.marginal = TRUE, method = "KMW")
> par (op)
> # Ieşire: Figura 7.22 supravieţuireBIV -1 .
```

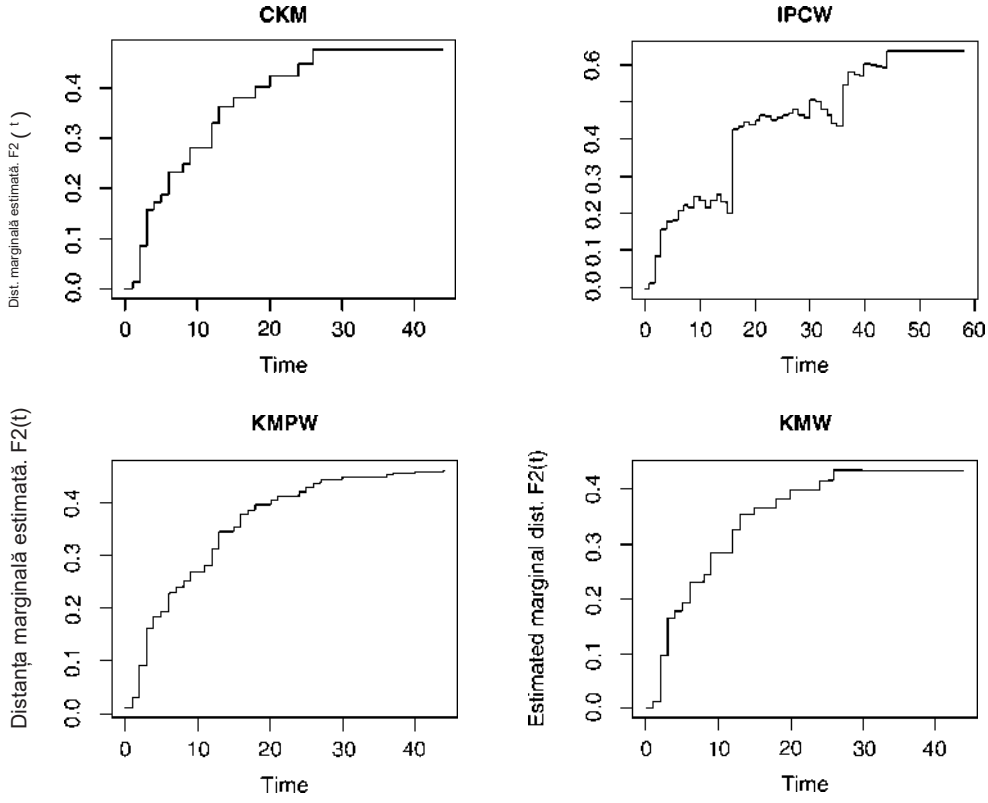


FIGURA 7.22
supravieţuireBIV -1.

```
> plot(bladderBIV_obj, plot.marginal = TRUE,
+ plot.bivariate = TRUE, method = "CKM")
```

Se aşteaptă confirmarea modificării paginii...

```
> # Ieşire: Figura 7.23 supravieţuireBIV-CKM -1 .
```

Se aşteaptă confirmarea modificării paginii...

```
> # Ieşire: Figura 7.24 supravieţuireBIV-CKM -2 .
```

Se aşteaptă confirmarea modificării paginii...

```
> # Ieşire: Figura 7.25 supravieţuireBIV-CKM -3 .

> plot(bladderBIV_obj, plot.bivariate = TRUE, method =
"IPCW")
```

Se așteaptă confirmarea modificării paginii...
> # *Ieșire :* Figura 7.26 SurvivalBIV-IPCW **-I** .

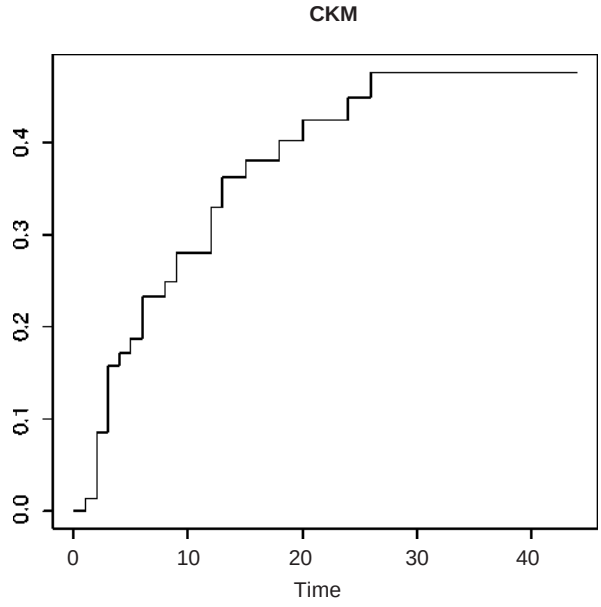


FIGURE 7.23 survivalBIV-CKM-1.

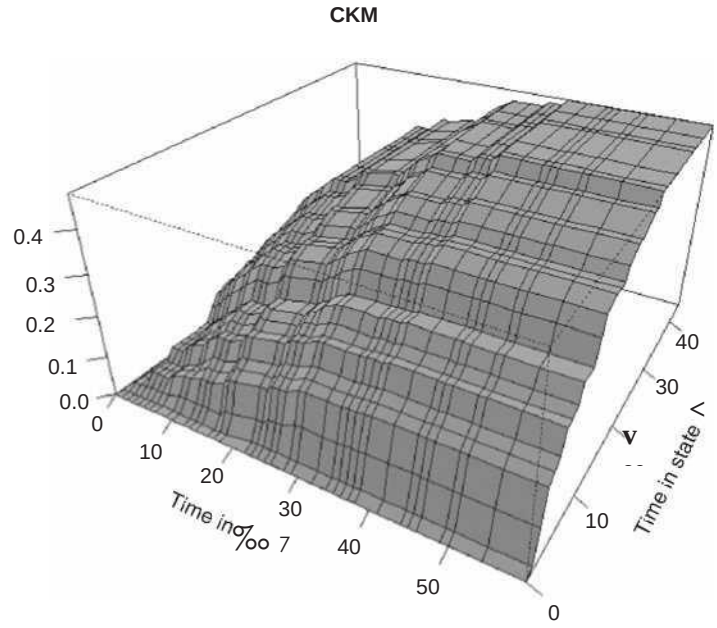


FIGURE 7.24 survivalBIV-CKM-2.

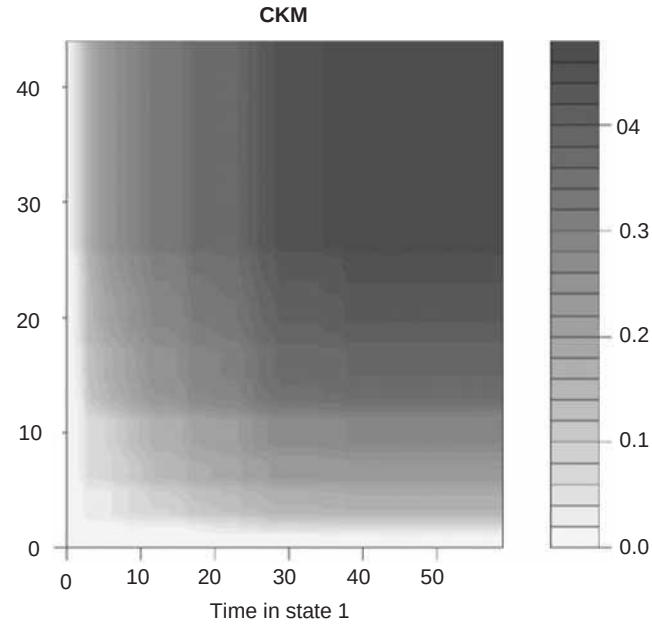


FIGURE 7.25 survivalBIV-CKM-3.

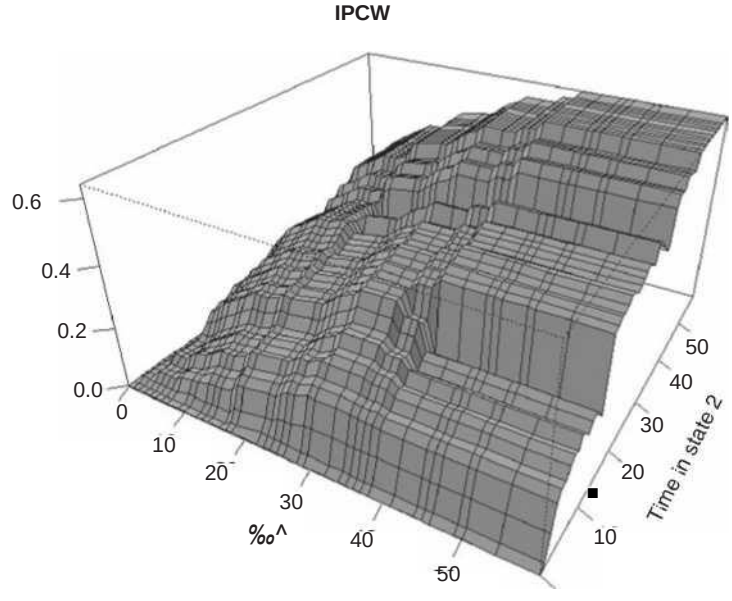


FIGURA 7.26 supraviețuireBIV-IPCW -1.

Se așteaptă confirmarea modificării paginii...
> # *Ieșire :* Figura 7.27 supraviețuireBIV-IPCW **-2** .

> plot(bladderBIV_obj, plot.persp = TRUE, method = "KMPW")

Se așteaptă confirmarea modificării paginii...

> *# Ieșire:* Figura 7.28 SurvivalBIV-KMPW .

> plot(bladderBIV_obj, plot.contour = TRUE, method = "KMW")

Se așteaptă confirmarea modificării paginii...
> *# Ieșire :* Figura 7.29 SupraviețuireBIV-KMW.

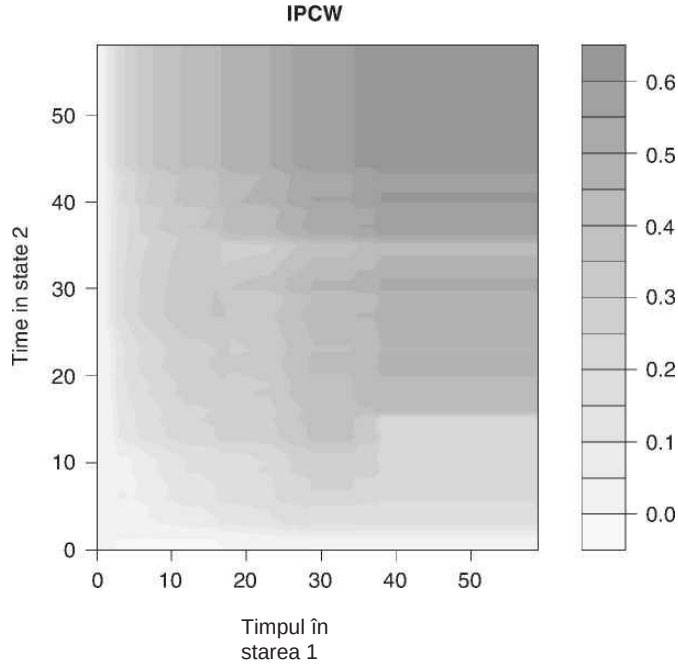


FIGURA 7.27 supraviețuireBIV-IPCW -2.

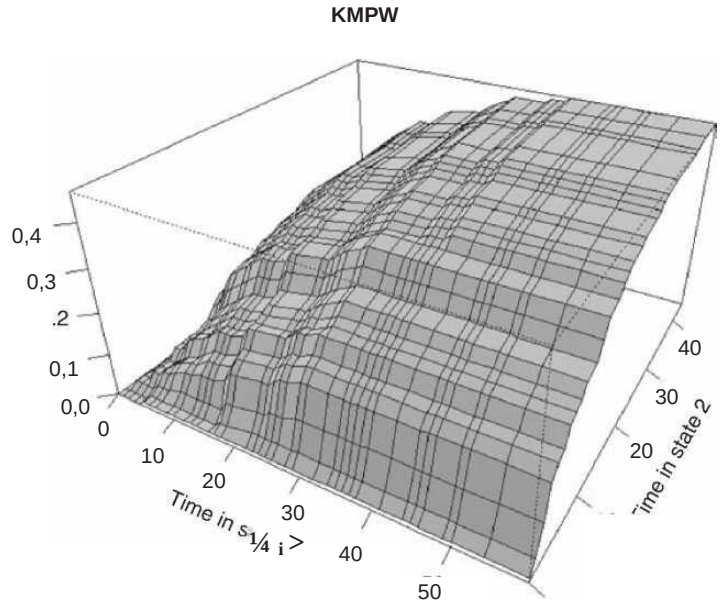


FIGURE 7.28 survivalBIV-KMPW.

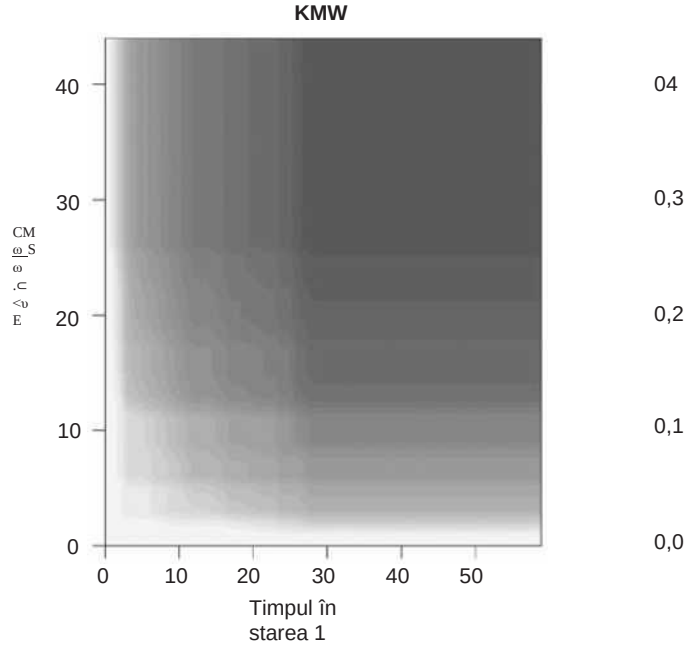


FIGURA 7.29 supraviețuireBIV-KMW.

7.4 REGRESII LOGISTICE

În biostatistică, precum și în statistică, unele rezultate iau forma unei **variabile dependente categorice** , care este o variabilă dependentă care poate lua un număr limitat de categorii, bazate pe una sau mai multe variabile predictoare.

O variabilă dependentă categorică **binomială** sau **binară** se referă la situația în care rezultatul observat poate avea doar două tipuri posibile, cum ar fi:

„Da" versus „Nu"
„Succes" versus „Eșuează"
„Bonavă" versus „Bine"
„Mort" versus „Viu"
„Ridicat" versus „Scăzut"
„Probabilitatea = 1" versus „Probabilitatea = 0"

Toate acestea pot fi reprezentate prin „1" versus „0".

În schimb, o variabilă dependentă categoric **multinomială poate avea trei sau mai multe tipuri posibile, cum ar fi:**

„Mai bine" versus „Neschimbat" versus „Mai rău"
„Ridicat" versus „Mediu" versus „Scăzut"
„Mărimea 4" versus „Mărimea 3" versus „Mărimea 2" versus „Mărimea 1"

Dacă probabilitatea relativă a fiecărui rezultat este x_i, atunci pentru posibilitatea I, pentru cazul general al *n* rezultate posibile:

$$x_1 + x_2 + x_3 + \ldots + x_n = 1$$

Pentru a modela variabilele dependente categoriale care au intervale limitate, o abordare este utilizarea unei **analize a modelului de regresie logistică (** Everitt &

Hothorn, 2006; Daniel, 2005; Dalgaard, 2002) . Capitolul 3 a prezentat analiza de regresie liniară care implică una, două sau mai multe variabile în Secțiunea 3.5.

Pentru un **sistem univariat** , modelul de regresie liniară simplă poate fi exprimat prin ecuația:

$$y = b_0 + b_{1x} + e \tag{7.12A}$$

în care

x este variabila *independentă* de valoare *X* ,
y este o valoare observată arbitrară a variabilei *dependente* continue *Y* , b_0 și b_1 sunt coeficienții regresiei liniare, iar *e* este termenul de eroare.

Într-o regresie liniară, când valoarea observată a lui *Y* este $m_{y|x}$ (media unei submulțimi de valori *Y* pentru o valoare dată a lui *X*), mărimea *e* (diferența dintre *Y observată* și linia de regresie) este zero . Ecuația (7.12) poate fi astfel scrisă după cum urmează:

$$m_{y|x} = b_0 + b_{1x} \tag{7.13A}$$

sau

$$E\,(y|x) = b_0 + b_{1x} \tag{7.14A}$$

În general, partea dreaptă a ecuațiilor (7.12A), (7.13A) și (7.14A) poate avea *orice valoare între infinitul negativ și infinitul pozitiv.*

Inadecvarea regresiei liniare simple atunci când *y este o* variabilă dependentă categorială

Când *y* este o variabilă dihotomică (o variabilă dependentă categorică simplă având doar două rezultate posibile; de exemplu, „1" sau „0"), modelul de regresie liniară simplă tocmai discutat *nu este* adecvat deoarece media (valoarea așteptată a lui *Y*) este probabilitatea ca *Y* = 1 și, prin urmare, este limitată la intervalul de la 0 la 1, inclusiv. Aceasta contrazice proprietatea împărtășită de ecuațiile (7.12A) , (7.13A) și (7.14A) ; și anume că partea dreaptă poate lua orice valoare.

Modelul de regresie logistică

În acest model, se poate lăsa *p* probabilitatea ca *Y* = 1:

$$p = P\,(Y = 1) \tag{7,15}$$

Raportul $p/(1-p)$ poate fi considerat după cum urmează:

(probabilitatea unui succes pentru *y*)/(1 - probabilitatea unui succes pentru *y*) = (probabilitatea unui succes pentru *y*)/(probabilitatea unui eșec pentru *y*) = cota, *p* , a răspunsului 0 luând valoarea 1.

Astfel, raportul $p/1-p$ poate lua valori între infinitul negativ ($-\infty$) și infinitul pozitiv ($+\infty$), așa cum este cerut de partea dreaptă a ecuațiilor (7.12) , (7.13) și (7.14) . Prin urmare, se poate rescrie ca

$$\log[\,p/(1-p)\,] = b_0 + b_{1x} \tag{7.16A}$$

Acesta este cunoscut sub numele de **modelul de regresie logistică** . Transformarea lui $m_{y|x}$ (adică, *p*) în $\log[\,p/(1-p)\,]$ este cunoscută ca **transformarea logit** . Ecuația (7.16A)

poate fi scrisă și ca

$$p = \exp(b_0 + b_1 x)/[1 + \exp(b_0 + b_1 x)], \quad (7.17A)$$

în care exp este inversul logaritmului natural $\log_e$ sau ln.

The Logit

Logit-ul probabilității p este logaritmul cotei π a răspunsului luând valoarea 1: logit (p)

$$= \log[p/(1-p)] \quad (7.18)$$

Acest model de regresie logistică este utilizat în mod obișnuit în biostatistică, epidemiologie și științele sănătății, în special atunci când este investigată o variabilă dependentă categorică . În analiza epidemiologică, se aplică în mod obișnuit în încercările de a modela probabilitatea (sau riscul) ca un subiect de caz să dobândească o anumită boală pe o anumită perioadă în care subiectul este expus la condiții cunoscute a fi asociate cu acea boală.

Pentru un **sistem multivariat** , ecuațiile rezultate corespunzătoare ecuațiilor (7.12A) , (7.13A) , (7.14A) , (7.16A) și (7.17A) sunt, respectiv:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_n x + e \quad (7.12B)$$

$$m_{y|x} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_n x_n \quad (7.13B)$$

$$E(y|x) = b_0 + b_1 x + b_2 x_2 + b_3 x_3 + \ldots + b_n x_n \quad (7.14B)$$

$$\log[p/(1-p)] = b_0 + b_1 x + b_2 x_2 + b_3 x_3 + \ldots + b_n x_n \quad (7.16B)$$

$$p = \exp(b_0 + b_1 x + b_2 x_2 + b_3 x_3 + \ldots + b_n x_n)/ [1 + \exp(b_0 + b_1 x + b_2 x_2 + b_3 x_3 + \ldots + b_n x_n] \quad (7.17B)$$

Analiza regresiei logistice

Această clasă de analiză este de obicei efectuată utilizând modele biostatistice aparținând clasei de modele liniare generalizate (vezi Capitolul 3 , Secțiunea 3.5, despre analiza datelor univariate, bivariate și multivariate). Aceste modele sunt, în general, caracterizate de următoarele elemente:

1. o distribuție de răspuns, cum ar fi distribuțiile binomiale și Poisson și
2. o funcție de legătură care transferă valoarea medie la o scară în care relația cu alte variabile poate fi exprimată ca aditivă și liniară. Pentru analiza de regresie logistică, funcția de legătură este de obicei logit p , așa cum este dată de ecuația (7.18) .

De exemplu, în analiza datelor de numărare, poate fi utilizat modelul Poisson multiplicativ, caz în care funcția de legătură este log l , unde l este media distribuției Poisson observate.

Toate aceste modele pot fi tratate de același algoritm, permițând analistului să definească funcții de legătură adecvate.

Modele liniare generalizate în R

Acest subiect a fost prezentat în Capitolul 3 , Secțiunea 3.5 (care a discutat despre analiza datelor univariate, bivariate și multivariate), unde s-a arătat că, în R , modelele liniare

generalizate sunt aplicate folosind funcția glm() , care este similară cu cea familiară funcția lm() . Aceste două funcții sunt, de asemenea, foarte asemănătoare în aplicare, prin aceea că ambele funcții folosesc aceleași formule de model și funcții de extracție (cum ar fi summary() etc.). Cu toate acestea, funcția glm() necesită un model liniar generalizat desemnat, care poate fi realizat folosind argumentul familiei din funcția glm() .

Următorul exemplu lucrat ilustrează modelarea practică de regresie logistică în aplicații biostatistice în epidemiologie și sănătate publică.

3. **Exemplul 7.5: Regresia** logistică pentru prezicerea unei variabile clinice cu valori binare folosind o legătură de regresie *binomială*

Pachetul de date Faraway conține 124 de seturi de rezultate ale investigațiilor clinice ale unui număr mare de subiecți de caz privind anumite sindroame metabolice (cum ar fi diabetul). În cadrul acestui pachet se află un set de date numit pima care conține înregistrările anumitor condiții de sănătate a 768 de subiecți cazați (fiecare incluzând diagnosticul de diabet) și indicele de masă corporală (IMC) corespunzător acestora .

(a) Folosind regresia logistică, arătați că un diagnostic de diabet la un subiect de caz este legat semnificativ de IMC-ul subiectului respectiv.
(b) Estimați probabilitatea ca cineva al cărui IMC este de 35 să aibă diabet.
(c) Pentru cineva cu probabilitatea percentilă de 95% de a avea diabet, estimați probabilitatea ca acea persoană să aibă un IMC de 35.

Solutii:

Următorul segment de cod R realizează calculele pentru acest exemplu. Mai întâi, apelați pachetul faraway și apoi examinați setul de date pima :

```
R 2.15.2 (26.10.2012) -- „Trick or Treat"
Copyright (C) 2012 Fundația R pentru Calcul Statistic
ISBN 3-900051-07-0
Platformă: i386-w64-mingw32/i386 (32 de biți)

R este un software gratuit și nu vine cu ABSOLUT NU GARANȚIE.
Sunteți binevenit să-l redistribuiți în anumite condiții.
Tastați „license()" sau „license()" pentru detalii despre distribuție.

Suport în limbaj natural, dar rulează într-o locație engleză
R este un proiect de colaborare cu mulți colaboratori.
Tastați „contributors()" pentru mai multe informații și „citation()" despre cum să
citați pachetele R sau R în publicații.

Tastați „demo()" pentru unele demonstrații, „help()" pentru ajutor on-line sau
„help.start()" pentru o interfață de browser HTML pentru a ajuta.
Tastați „q()" pentru a părăsi R .

[Spațiul de lucru salvat anterior a fost restaurat]
```

Mai întâi, examinați fișierul de date investigat:

```
>
> install.packages(„departe")
Instalarea pachetelor în „C:/Users/bertchan/Documents/R/win-library/2.15"
(deoarece „lib" nu este specificat)
> -- Vă rugăm să selectați o oglindă CRAN pentru a fi utilizată în această
sesiune --
```

se încearcă URL-ul „
http://cran.stat.ucla.edu/bin/windows/contrib/2.15/faraway
_1.0.5.zip'

Tip de conținut „aplicație/zip” lungime 593494 octeți (579 Kb)
URL deschis
descărcat 579 Kb

pachetul „faraway” a fost despachetat cu succes și sumele MD5 verificate

Pachetele binare descărcate sunt în
C:\Users\bertchan\AppData\Local\Temp\RtmpwPV84x\downloaded
_packages

```
> bibliotecă (departe)
> ls("pachet:departe")
```

[1] „aatemp”	**„abraziune”**	**„aflatoxină”**	**"Africa"**	**"lucernă"**
[6] „amlxray”	**„hrană**	**„gândacă”**	**"fericire"**	**"rupere"**
[11] „broccoli”	**pentru**	**"Chicago"**	**"chiczip"**	**"chimiss"**
	copii”			
[71] „nes96”	**"oatvar"**	**"miros"**	**"Ohio"**	**"orings"**
[76] „ozon”	**"parstum"**	**"arahide"**	**"penicilină**	**"pima"**
[81] „conductă”	**"pneumo"**	**"potuse"**	**"prostata"**	**"prplot"**
[121] „wavesolder”	**„wbca”**	**„weldstrength”**	**„grâu”**	

```
> pima
```

	gravidă	**Glucoză**	**diastolică**	**triceps**	**insulina**	**bmi**	**test de vârstă**	**pentru diabet**
16		**148**	**72**	**35**	**0**	**33,6**	**0,627**	**50**	**1**
21		**85**	**66**	**29**	**0**	**26,6**	**0,351**	**31**	**0**
38		**183**	**64**	**0**	**0**	**23.3**	**0,672**	**32**	**1**
767	**1**	**126**	**60**	**0**	**0**	**30.1**	**0,349**	**47**	**1**
768	**1**	**93**	**70**	**31**	**0**	**30,4**	**0,315**	**23**	**0**

```
> pima$test
```

```
[1]   1 0 1 0 1 0 1 0 1 1 0 1 0 1 1 1 1 1 0 1 0 0 1 1 1 1 1 0 0 0 0 1 0 0 0 0 0
[38]  1 1 1 0 0 0 1 0 1 0 0 1 0 0 0 0 1 0 0 1 0 0 0 0 1 0 0 1 0 1 0 0 0 1 0 1 0
[75]  0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 1
[112] 1 0 0 1 1 1 0 0 0 1 0 0 0 1 1 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
[149] 0 0 0 0 1 0 1 1 0 0 0 1 0 0 0 0 1 1 0 0 0 0 1 1 0 0 0 1 0 1 0 1 0 0 0 0 0
[186] 1 1 1 1 1 0 0 1 1 0 1 0 1 1 1 0 0 0 0 0 0 1 1 0 1 0 0 0 1 1 1 1 0 1 1 1 1
[223] 0 0 0 0 0 1 0 0 1 1 0 0 0 1 1 1 1 0 0 0 1 1 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0
[260] 1 0 1 0 0 1 0 1 0 0 1 1 0 0 0 0 0 1 0 0 0 1 0 0 1 1 0 0 1 0 0 0 1 1 1 0 0
[297] 1 0 1 0 1 1 0 1 0 0 1 0 1 1 0 0 1 0 1 0 0 1 0 1 0 1 1 1 0 0 1 0 1 0 0 0 1
[334] 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 1 1 0 1 1 0 0 1 0 0 1 0 0 1
[371] 1 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 1 1 1 0 0 1 0 0 1 0 0 1 0 1 1 0 1 0 1 0 1
[408] 0 1 1 0 0 0 0 1 1 0 1 0 1 0 0 0 0 1 1 0 1 0 1 0 0 0 0 0 1 0 0 0 0 1 0 0 1
[445] 1 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1
[482] 0 0 0 1 1 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 1 1 0
[519] 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 1 1 1 0 0 1 1 0 0 0 0 0 0 0 0
[556] 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 1 0 1 0 1 0 1 0
[593] 1 0 0 1 0 0 1 0 0 0 0 0 1 1 0 1 0 0 0 0 0 1 1 0 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0
[630] 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 1 0 0 0 1 1 1 0 0 0 0 0 0 0 1 0 0 0 1 0 1 1 1 1 0
[667] 1 1 0 0 0 0 0 0 0 0 1 1 0 1 0 0 1 0 1 0 0 0 0 0 0 1 0 1 0 1 0 1 1 0 0 0 0 1 1
[704] 0 0 0 1 0 1 1 0 0 1 0 0 1 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 1
[741] 1 0 0 1 0 0 1 0 1 1 1 0 0 1 1 1 0 1 0 1 0 1 0 0 0 0 0 1 0
```

(a) Corelarea setului de date pima printr-o funcție de regresie logistică lr() , cu bmi ca variabilă și folosind un model liniar generalizat glm() cu o funcție binomială binomial () ca funcție de legătură:

```
> lr <- glm(f ~ bmi, family = binom, data = pima)
```

Pentru a examina rezultatele critice calculate, apelați:

```
> rezumat (lr)
> # Ieșire :
```

Apel:
glm(formula = f ~ bmi, familie = binom, date = pima)

Reziduuri de devianta:

Min 1Q Median 3Q Max
-1,9209 -0,9178 -0,6838 1,2351 2,7244

Coeficienți:

Estimare Std. Valoarea z eroare Pr(>|z|) (Interceptare) -
3,68641 0,40896 -9,014 < 2e-16 * bmi 0,09353 0,01205**
7,761 8,45e-15 *

Signif. coduri: 0 '*' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 " 1**

(Parametrul de dispersie pentru familia binomială considerat a fi 1)

Devianța nulă: 993,48 pe 767 grade de libertate
Devianța reziduală: 920,71 pe 766 grade de libertate
AIC: 924,71

Numărul de iterații Fisher Scoring: 4

În mod clar, variabila IMC din setul de date este semnificativă.

(b) Pentru a estima probabilitatea ca cineva al cărui IMC este 35 să aibă diabet, puteți utiliza modelul de regresie logistică pentru a calcula probabilitatea ca un subiect cu un IMC mediu de 35 să fie pozitiv pentru un diagnostic de diabet:

```
> predict(lr, type = „răspuns", date noi = date noi)
> # Ieșire:
   1
0,39823
```

Astfel, conform acestui model, probabilitatea calculată este de 39,8%, sau aproape 40%, sau 2 din 5.

(c) Pentru cineva cu probabilitatea percentila de 95% de a suferi de diabet, estimați probabilitatea acestui caz; adică pentru un subiect de caz al cărui IMC este 35. > newdata1 <- data.frame(bmi = 35)

```
> predict(lr, type = „răspuns", date noi = date noi1)
> # Ieșire:
  95%
0,6144066
```

Astfel, conform probabilității calculate de acest model, un subiect de caz cu un IMC de 35 ar avea o probabilitate de 61,4% (sau aproximativ 2/3, sau 2 din 3) de a se afla în percentila 95 dintre cei diagnosticați de diabet. .

Întrebări de revizuire pentru Secțiunea 7.4

1. În regresia logistică, este recomandabil să folosiți modele aditive pentru probabilități? De ce sau de ce nu?
2. În modelarea regresiei logistice, nu există termeni de eroare și nici parametri de varianță ca în distribuția normală. De ce?
3. În regresia logistică, parametrii modelului pot fi estimați prin metoda probabilității maxime? De ce sau de ce nu?
4. În regresia logistică, funcția de legătură, care transferă valoarea medie la o scară în care relațiile cu alte variabile pot fi descrise, este liniară și aditivă. Arătați că acest lucru poate fi realizat folosind următoarea funcție de legătură:

$$\log p = \log \{p/(1 - p)\}$$

5. În modelele liniare generalizate, funcția predict() poate fi utilizată pentru a stabili probabilitatea de apariție a anumitor evenimente. Având în vedere un set de date de rezultate observate într-un set de condiții de variabile independente, schițați o procedură pentru utilizarea funcției predict() pentru a stabili o curbă de probabilitate prezisă pentru oricare dintre variabilele independente date.
6. (a) Când ar trebui utilizat modelul de regresie logistică?
   (b) Definiți termenii *odds* și *odds ratio* . Dați exemple pentru fiecare.
7. Sugerați un exemplu în epidemiologie în care analiza de regresie logistică este adecvată :
   (a) când variabila independentă este continuă
   (b) când variabila independentă este discretă
8. Dacă p = probabilitate și o = cotă, atunci

$$o = [\, p /(1 - p \,)]$$

   si astfel

$$p = [\, o /(1+ o \,)]$$

   (a) Care sunt limitele superioare și inferioare pentru p ? Pentru o ?
   (b) Ce avantaje are folosirea lui p ? De ce?
   (c) Ce avantaje are utilizarea lui o ? De ce?

9. Într-un sistem cu n covariate, modelul de regresie logistică corespunzător poate fi exprimat ca

$$\log(\text{cote}) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_n x_n + e \,.$$

   b_1 poate fi considerată după cum urmează: O creștere unitară în x_1, cu x_2, x_3, …, x_n menținut fix, crește șansele de succes log cu b_1 sau crește șansele de succes cu un factor de exp(b_1).

   Scrieți o expresie pentru „cote” atunci când dependențele de covariate sunt neliniare.

10. Pentru un rezultat dat, dacă probabilitatea de
    (a) „succesul” în prezența Condiției A este p_1 și
    (b) „eșecul” în absența condiției A este p_2,

    care este sensul raportului p_1/p_2? Explicați și dați exemple.

Exerciții pentru Secțiunea 7.4

1. Modelele de regresie logistică sunt adesea folosite pentru a calcula o curbă de supraviețuire pentru datele cenzurate longitudinale. Următorul segment de cod R calculează o estimare a unei curbe de supraviețuire pentru datele cenzurate utilizând fie metoda K–M, fie metoda Fleming–Harrington pentru calcularea funcției de supraviețuire prezisă. Pentru datele privind riscurile concurente, calculează curba de incidență cumulativă. Aceasta se numește survfit-ul pachetului de supraviețuire . funcție de formulă cu o valoare implicită diferită pentru conf.type (bază log-log). Mai mult, atributele variabilei timp eveniment sunt salvate.

   Utilizare:
   survfit(formulă, date, ...)

   Argumente:

   formula Un obiect formulă, care trebuie să aibă un obiect Surv ca răspuns în stânga operatorului ~ și, dacă este necesar, termeni separați prin operatorii + în partea dreaptă. Unul dintre termeni poate fi un obiect de strat . Pentru o singură curbă de supraviețuire, partea dreaptă ar trebui să fie ~ 1 .
   date Un cadru de date în care se interpretează variabilele numite în argumentele formulei, subsetului și ponderilor .
   ... Vezi survfit.formula
   Detalii Consultați survfit.formula pentru detalii.
   Valoare Un obiect din clasa "survfit" . Consultați survfit.object pentru detalii. Metodele definite pentru obiectele survfit sunt tipărirea , trasarea , liniile și punctele .

   Vezi și:
   survfit.cph pentru curbele de supraviețuire din modelele Cox; print , plot , lines , coxph , Surv , strata.

   Autori:
   Thomas Lumley (tlumley@u.washington.edu) ; și Terry Therneau

   # **Exemple:**

```
cere (supraviețuire)
#   fi ta curba Kaplan-Meier și trasează-o:
fit <- survfit(Surv(time, status) ~ x, data = aml)
plot(potrivire, lty = 2:3)
legenda(100, .8, c(„Menținut”, „Neîntreținut”), lty = 2:3)
```

```
#   fi ta Cox hazarde proporționale modelați și reprezentați grafic
#   supraviețuirea prezisă pentru un bărbat de 60 de ani:
fit <- coxph(Surv(futime, fustat) ~ age, data = ovarian) plot(survfit(fit,
newdata=data.frame(age=60)),
      xscale=365.25, xlab = „Ani", ylab="Survival")

#   Iată setul de date de la Turnbull.
#   Nu există subiecți cenzurați pe interval, doar cenzurați în stânga (status=3), #
cenzurați în dreapta (starea 0) și evenimente observate (statusul 1).

Timp
1  2 3 4
Tipul de observație
moartea 12 6 2 3
pierderi 3 2 0 3
intrare târzie 2 4 2 5

tdata <- data.frame(time =c(1,1,1,2,2,2,3,3,3,4,4,4), status=rep(c(1,0,2),4 ), n
                          =c(12,3,2,6,2,4,2,0,2,3,3,5))
fit <- survfit(Surv(time, time, status, type='interval') ~1, data=tdata, weights=n)

#   Timp până la progresie/deces pentru pacienții cu gammapatie monoclonală:
Curbe de risc concurente (incidență cumulativă)
fit1 <- survfit(Surv(stop, event=='progresie') ~1, data=mgus1, subset=(start==0))
fit2 <- survfit(Surv(stop, status) ~1, data=mgus1, subset=(start==0), etype=event)
#   riscuri concurente;
#   Curbele CI sunt întotdeauna reprezentate de la 0 în sus, nu de la 1 în jos: plot(fit2,
fun='event', xscale=365.25, xmax=7300,
            mark.time=FALSE,col=2:3, xlab="Ani după diagnosticarea MGUS")
linii (fit1, fun='eveniment', xscale=365.25, xmax=7300,
            mark.time=FALSE,conf.int=FALSE)
text(10, .4, „Risc concurent: moarte", col=3)
text(16, .15,"Risc de concurență: progresie", col=2)
text(15, .30,"KM:prog")
```

Următorul segment de cod R preia executarea acestui exercițiu:

```
>
> install.packages("rms")
```

Instalarea pachetelor în „C:/Users/bertchan/Documents/R/win-library/2.14"
(deoarece „lib" nu este specificat)
--- Vă rugăm să selectați o oglindă CRAN pentru a fi utilizată în această sesiune --
se încearcă URL-ul „
http://cran.stat.ucla.edu/bin/windows/contrib/2.14/rms_3.6-3.zip"
Tip de conținut „aplicație/zip" lungime 933733 octeți (911 Kb)
URL deschis
descărcat 911 Kb

pachetul „rms" a fost despachetat cu succes și sumele MD5 au fost verificate

```
Pachetele descărcate sunt în
C:\Users\bertchan\AppData\Local\Temp\RtmpyAfTbf\downloaded
_packages
> bibliotecă (rms)
Se încarcă pachetul necesar: Hmisc
Se încarcă pachetul necesar: supraviețuire
Se încarcă pachetul necesar: spline
Biblioteca Hmisc de Frank E Harrell Jr
Tastați library(help='Hmisc'), ?Overview sau ?Hmisc.Overview') pentru a vedea
documentația generală.
NOTĂ: Hmisc nu mai redefinește [.factor pentru a reduce nivelurile neutilizate
la subsetare. Pentru a obține vechiul comportament al Hmisc, tastați
dropUnusedLevels().
Pachetul atașat: „Hmisc”
Următoarele obiecte sunt mascate din „pachet:supraviețuire”:
   untangle.specials
Următoarele obiecte sunt mascate din „pachet:bază”:
   format.pval, rotund.POSIXt, trunc.POSIXt, unități
Se atașează pachetul: „rms”
Următoarele obiecte sunt mascate din „pachet:supraviețuire”:
   Surv
> ls(„pachet:rms”)
[1]     „%ia%”
[2]     „.rms”
[3]     „.Surv”
[4]     „AIC.rms”
[5]     "anova.rms"
[205]   "Supraviețuire"
```

```
>
>  cere (supraviețuire)
>  # fi ta Kaplan-Meier și complotează-l:
>  fit <- survfit(Surv(time, status) ~ x, data = aml)
>  plot(potrivire, lty = 2:3)
> legendă(100, .8, c(„Menținut", „Neîntreținut"), lty = 2:3)
>  # Ieșire: Figura 7.30 .
```

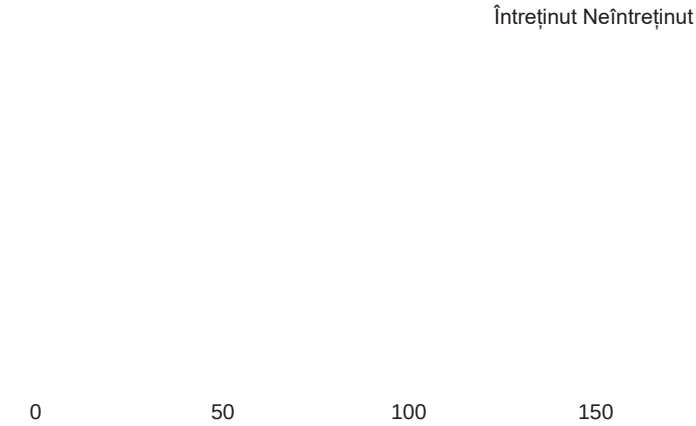


FIGURA 7.30 Regresia logistică, diagrama de supraviețuire K–M.

(a) Observați ordonata (axa verticală) a diagramei de supraviețuire. În ce fel seamănă această diagramă cu o diagramă de probabilitate?

(b) Observați abscisa (axa orizontală) a diagramei de supraviețuire. Care sunt unitățile pentru această axă?

(c) Ce alte caracteristici ale acestui diagramă de supraviețuire sunt caracteristice unui grafic tipic de regresie logistică?

(d) Se intersectează cele două grafice de supraviețuire de pe diagrama de supraviețuire? De ce sau de ce nu?

(e) În diagrama de supraviețuire, diagrama „Menținută" este întotdeauna deasupra diagramei „Neîntreținute". De ce?

2. *Validarea probabilităților prezise folosind regresia logistică: Funcția* val. problema()

Funcția val.prob() poate fi utilizată pentru validarea probabilităților prezise în raport cu evenimente binare. Având în vedere un set de probabilități prezise p sau logit -odds prezis și un vector de rezultate binare y care nu au fost utilizate în dezvoltarea predicțiilor p sau logit , funcția val.prob() calculează următoarele statistici și indici:

(1) Corelația rangului $D_{xy\ între}$ p și y [2(C – 0,5), C = zona ROC]. În compararea și analiza curbelor, ROC este o măsurătoare pentru compararea performanței de diagnosticare a două sau mai multe teste de diagnosticare. Performanța de

diagnosticare a unui test, sau acuratețea unui test, este capacitatea sa de a discrimina cazurile bolnave de cazurile normale. Când se iau în considerare rezultatele unui anumit test în două populații , o populație cu o boală și cealaltă populație fără boală, rareori se va observa o separare perfectă între cele două grupuri; distribuția rezultatelor testelor se va suprapune cel mai probabil.

(2) indicele R pătrat.
(3) Indicele de discriminare D (model logistic).
(4) Indicele de incredere U.
(5) χ^2 cu 2 grade de libertate, și valoarea lui *p* .
(6) Indicele de calitate Q.
(7) Scorul Brier (diferența medie pătrată în p și y), interceptarea și panta.
(8) E_{max} = diferența maximă absolută a probabilităților estimate și calibrate.
(9) Spiegelhalter *Z* pentru precizia calibrării și valoarea *p cu două cozi* . Dacă pl = TRUE , programul trasează o curbă de calibrare logistică adaptată și, opțional, o potrivire neparametrică lină folosind proporții mai mici (p,y,iter=0) și grupate față de probabilitatea medie prezisă în grup. Dacă probabilitățile sau logit-urile prezise sunt constante, statisticile sunt returnate și nu se produce niciun grafic.

Exemple:

```
# Potriviți modelul logistic pe 100 de observații simulate din
# model real dat de
# Prob(Y=1 dat X1, X2, X3) = 1/(1+exp[-(-1 + 2X1)]),
# unde X1 este o variabilă uniformă aleatorie [0,1]. Prin urmare, X2 și X3
# sunt irelevante. După ajustarea unui model aditiv liniar în X1, X2,
# și X3, coeficienții sunt utilizați pentru a prezice Prob(Y=1) pe a
# eșantion separat de 100 de observații. Rețineți că împărțirea datelor este
# o metodă de validare ineficientă, cu excepția cazului în care n > 20.000.
set.seed(1)
n <- 200
x1 <- runif(n)
x2 <- runif(n)
x3 <- runif(n) logit <- 2*(x1-.5) P <- 1/(1+exp(-logit)) y <- ifelse(runif(n)<=P, 1, 0) d <-
data.frame(x1,x2,x3,y) f <- lrm(y ~ x1 + x2 + x3, subset=1:100) pred.logit <- predict(f,
d[101:200,]) phat <- 1/(1+exp(-pred.logit)) val.prob(phat, y[101:200], m=20, cex=.5) #
subgrupe de 20 obs.
# Validați predicțiile mai strict prin stratificarea dacă
# x1 este deasupra sau sub mediana:

v <- val.prob(phat, y[101:200], group=x1[101:200], g.group=2)
v
complot (v)
plot(v, flag=funcție(statistici) ifelse(
statistici[,'ChiSq2'] > qchisq(.95,2) |
statistici[,'B ChiSq'] > qchisq(.95,1), '*', '') )
# Începe rândurile de statistici în diagramă corespunzătoare semnificației
# calibrare greșită la nivelul 0,05 în loc de cea implicită, 0,01: plot(val.prob(phat,
y[101:200], group=x1[101:200], g.group=2), col=1:3) # 3 culori (1 pentru salopeta)
```

```
#   Curbe de calibrare ponderate
#   plot(val.prob(pred, y, group=varsta, greutate=frecvențe))

#   Potriviți modelul logistic pe 100 de observații simulate din
#   model real dat de
#   Prob(Y=1 dat X1, X2, X3) = 1/(1+exp[-(-1 + 2X1)]),
#   unde X1 este o variabilă uniformă aleatorie [0,1]. Prin urmare, X2 și X3
#   sunt irelevante. După ajustarea unui model aditiv liniar în X1, X2,
#   și X3, coeficienții sunt utilizați pentru a prezice Prob(Y=1) pe a
#   eșantion separat de 100 de observații. Rețineți că împărțirea datelor
#   este o metodă de validare ineficientă, cu excepția cazului în care n > 20.000.
set.seed(1)
n <- 200
x1 <- runif(n)
x2 <- runif(n)
x3 <- runif(n)
logit <- 2*(x1-.5)
P <- 1/(1+exp(-logit))
y <- ifelse(runif(n)<=P, 1, 0)
d <- data.frame(x1,x2,x3,y)
f <- lrm(y ~ x1 + x2 + x3, subset=1:100)
pred.logit <- predict(f, d[101:200,])
phat <- 1/(1+exp(-pred.logit))
val.prob(phat, y[101:200], m=20, cex=.5) # subgrupe de 20 obs.
```

Următorul segment de cod R preia executarea acestui exercițiu:

```
#   Potriviți modelul logistic pe 100 de observații simulate din
#   model real dat de
#   Prob(Y=1 dat X1, X2, X3) = 1/(1+exp[-(-1 + 2X1)]),
#   unde X1 este o variabilă uniformă aleatorie [0,1]. Prin urmare, X2 și X3
#   sunt irelevante. După ajustarea unui model aditiv liniar în X1, X2,
#   și X3, coeficienții sunt utilizați pentru a prezice Prob(Y=1) pe a
#   eșantion separat de 100 de observații. Rețineți că împărțirea datelor este
#   o metodă de validare ineficientă, cu excepția cazului în care n > 20.000.

    set.seed(1)
    n <- 200
    x1 <- runif(n)
    x2 <- runif(n)
    x3 <- runif(n)
    logit <- 2*(x1-.5)
    P <- 1/(1+exp(-logit))
    y <- ifelse(runif(n)<=P, 1, 0)
    d <- data.frame(x1,x2,x3,y)
    f <- lrm(y ~ x1 + x2 + x3, subset=1:100)
    pred.logit <- predict(f, d[101:200,])
```

```
phat <- 1/(1+exp(-pred.logit))
val.prob(phat, y[101:200], m=20, cex=.5) # subgrupe de 20 obs. >
>  Validați predicțiile mai strict prin stratificarea dacă
>  x1 este deasupra sau sub mediana:
v <- val.prob(phat, y[101:200], group=x1[101:200], g.group=2)
v
plot(val.prob(phat, y[101:200], group=x1[101:200], g.group=2),
col=1:3) # 3 culori (1 pentru general)
>  Curbe de calibrare ponderate:
>  plot(val.prob(pred, y, grup=vârstă, greutăți=frecvențe)) val.
>
>  # Potriviți modelul logistic pe 100 de observații simulate din
>  # model real dat de Prob(Y=1 dat X1, X2, X3) =
>  # 1/(1+exp[-(-1 + 2X1)]),
>  # unde X1 este o variabilă uniformă aleatorie [0,1]. Prin urmare, X2 și
>  # X3 sunt irelevante. După ajustarea unui model aditiv liniar în X1,
>  # X2 și X3, coeficienții sunt utilizați pentru a prezice Prob(Y=1) .
>  # un eșantion separat de 100 de observații. Rețineți că datele
>  # împărțirea este o metodă de validare ineficientă, cu excepția cazului în care n >
20.000.
>  set.seed(1)
>  n <- 200
x1 <- runif(n)
>  x2 <- runif(n)
>  x3 <- runif(n)
>  logit <- 2*(x1-.5)
>  P <- 1/(1+exp(-logit))
>  y <- ifelse(runif(n)<=P, 1, 0)
>  d <- data.frame(x1,x2,x3,y)
>  f <- lrm(y ~ x1 + x2 + x3, subset=1:100)
>  pred.logit <- predict(f, d[101:200,])
>  phat <- 1/(1+exp(-pred.logit))
>  val.prob(phat, y[101:200], m=20, cex=.5) # subgrupuri din 20
>  obs
>  # Ieșire:
```

Dxy	C (ROC)	R2	DD:Chi-sq	D:p
0,320528211	0,660264106	0,094994525	0,063901031	7,390103054	0,006558373

UU:Chi-sq	U:p	Q	Brier	Intercept
0,019128530	0,087146981	0,957362194	0,083029561	0,231896346	0,052287214

Pantă	Emax	S:z	S:p	Eavg
0,956517813	0,019247505	0,144009939	0,885492622	0,025838218

```
>  # Validați predicțiile mai riguros prin stratificarea
>  # dacă x1 este deasupra sau sub mediană:
>  v <- val.prob(phat, y[101:200], grup=x1[101:200],
```

```
+                  g.grup=2)
> v
>  # Ieşire:
```

	n	Pavg	Obs	ChiSq	ChiSq2	eavg	Eavg/P90	Med	C	B
[0,0131,0,526	50	0,364	0,36	0,004	0,149	0,046	0,170	1.228	0,637	0,220
[0,5260,0,993	50	0,590	0,62	0,195	1.902	0,058	0,213	1.083	0,514	0,244
În general	100	0,477	0,49	0,073	0,088	0,026	0,065	1.084	0,660	0,232

	B	B cal
[0,0131,0,526	0,073	0,212
[0,5260,0,993	0,327	0,230
În general	0,021	0,230

Quantile de probabilităţi prezise

	0,01	0,025	0,05	0,1	0,25	0,5	0,75	0,9	0,95	0,975
[0,0131,0,526)	0,213	0,222	0,242	0,270	0,301	0,364	0,407	0,471	0,516	0,531
[0,5260,0,993]	0,397	0,411	0,444	0,488	0,535	0,601	0,650	0,668	0,715	0,748
În general	0,216	0,243	0,271	0,292	0,365	0,479	0,598	0,654	0,668	0,714

	0,99
[0,0131,0,526)	0,545
[0,5260,0,993]	0,764
În general	0,750

```
>  complot (v)
>  # Ieşire:
```

Grupa [0,0131,0,526) [0,5260,0,993] Total
n 50 50 100
Pavg 0,364 0,590 0,477
Obs 0,36 0,62 0,49
ChiSq 0,0 0,2 0,1

ChiSq2 0,1 1,9 0,1
Mediu 0,046 0,058 0,026
Eavg/P90 0,170 0,213 0,065
Med OR 1.228 1.083 1.084
C 0,637 0,514 0,660
B 0,220 0,244 0,232
B ChiSq 0,1 0,3 0,0
B cal 0,212 0,230 0,230

```
> # Ieşire: Figura 7.31
> # Stelele rânduri de statistici în diagramă corespunzătoare semnificantului
> # calibrare greşită la nivelul 0,05 în loc de cel implicit, 0,01:
> plot(val.prob(phat, y[101:200], group=x1[101:200],
>                         g.grup=2),
+                         col=1:3) # 3 culori (1 pentru general)
```

Group	n	Pavg	Obs	ChiSq	ChiSq2	Eavg	Eavg/P90	Med OR	C	B	B Ch	
[0.0131, 0.526]	50	0.364	0.36	0.0	0.1	0.046	0.170	1.228	0.637	0.220	Z	0.1
[0.5260, 0.993]	50	0.590	0.62	0.2	1.9	0.058	0.213	1.083	0.514	0.244²		0.3
Overall	100	0.477	0.49	0.1	0.1	0.028	0.085	1.084	0.660	0.232		0.0

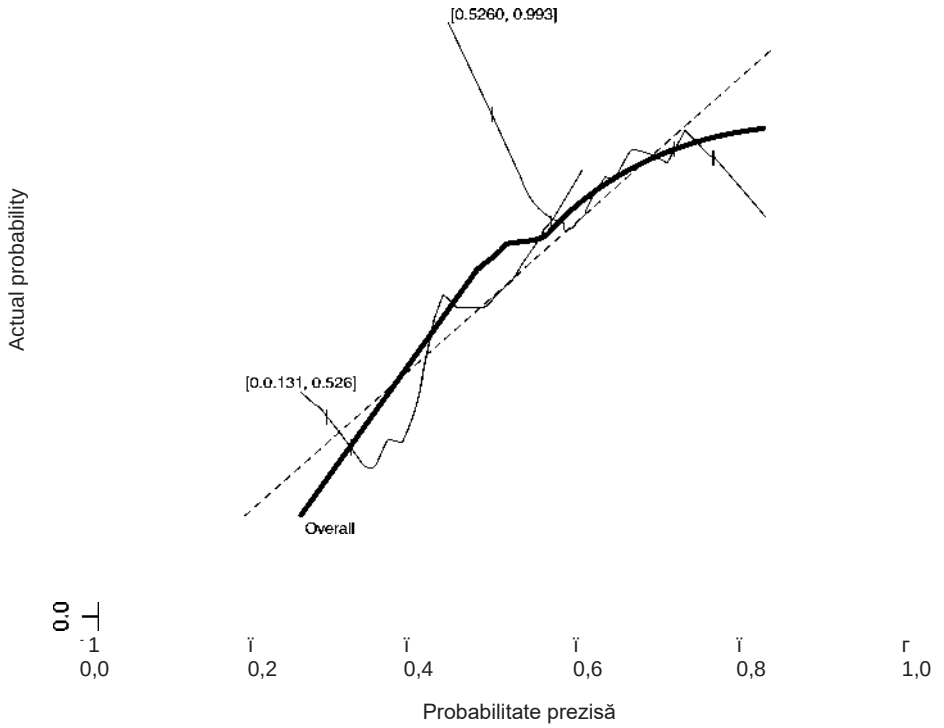


FIGURA 7.31 Regresia logistică: Validarea probabilităților prezise.

```
> # Ieşire:
```

Grupa [0,0131,0,526) [0,5260,0,993] Total n 50 50 100

Pavg 0,364 0,590 0,477
Obs 0,36 0,62 0,49
ChiSq 0,0 0,2 0,1
ChiSq2 0,1 1,9 0,1
Mediu 0,046 0,058 0,026
Eavg/P90 0,170 0,213 0,065
Med OR 1.228 1.083 1.084
C 0,637 0,514 0,660
B 0,220 0,244 0,232
B ChiSq 0,1 0,3 0,0
B cal 0,212 0,230 0,230

> # *Ieşire:* Figura 7.32 .

Grup	n	Pavg	Obs	ChiSq	ChiSq2	Eavg	Eavg/P90	Med	SAU	CBB	Ch
[0,0131, 0,526]	50	0,364	0,36	0,0	0,1	0,046	0,170	1,228	0,637	0,220	Z 0,1
[0,5260, 0,993]	50	0,590	0,62	0,2	1,9	0,058	0,213	1,083	0,514	0,244	² 0,3
Total	100	0,477	0,49	0,1	0,1	0,028	0,085	1,084	0,660	0,232	0,0 Z

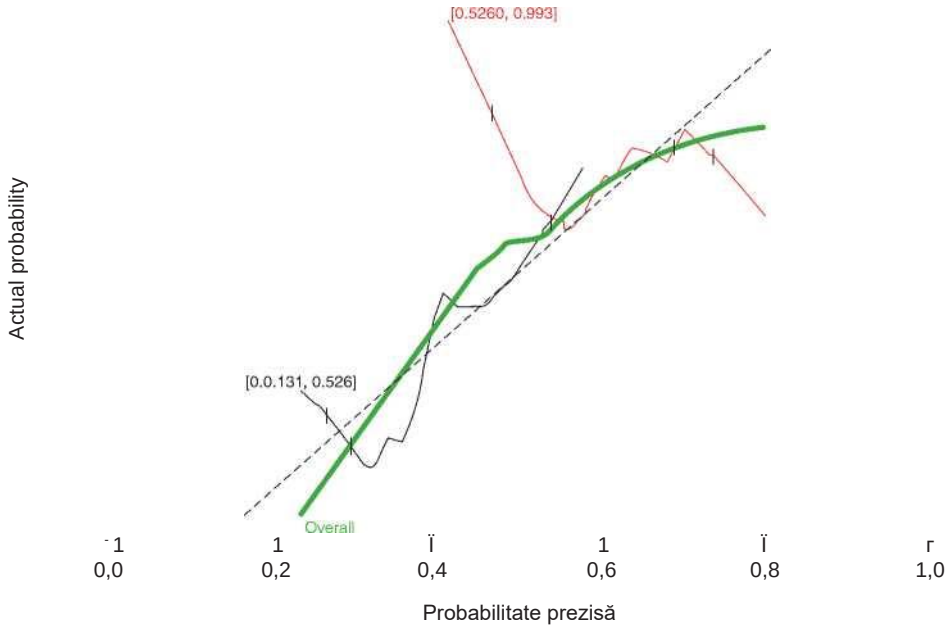


FIGURA 7.32 Regresia logistică: Validarea probabilităților prezise.

(a) Declarația acestui exercițiu a indicat că opt criterii sunt calculate prin această metodologie de utilizare a regresiei logistice pentru a valida probabilitățile prezise. Examinați fiecare dintre aceste opt criterii (începeți prin a căuta pe Internet „ corelația rangului D_{xy}”).

(b) Clasează aceste opt criterii, începând cu cel mai eficient criteriu și explică motivele pentru clasarea ta.

(c) În figurile 7.30 și 7.31 , dacă probabilitățile prezise ar fi aceleași cu probabilitățile reale, atunci punctele/liniile de date prezise ar fi toate pe linia diagonală (45 de

grade). Comentați acuratețea probabilităților estimate calculate .

(d) În ambele parcele, se pare că predicțiile sunt mai precise în jurul secțiunii din mijloc, unde probabilitatea este de aproximativ 0,5. Comentați această caracteristică a modelului de regresie logistică ca instrument predictiv.

(e) Oferiți o justificare pentru aplicarea modelului de regresie logistică ca instrument statistic pentru predicția succeselor (și eșecurilor) în problemele de probabilitate aplicate în biostatistică și epidemiologie.

3. *Estimarea cvasi-probabilității.*
În ajustarea modelelor generalizate cu distribuții de eroare Poisson sau binomiale, este posibil să se detecteze supradispersia prin compararea devianței reziduale cu gradele sale de libertate. Cele două cantități ar trebui să fie aproximativ egale pentru un model care se potrivește bine. Atunci când devianța este mult mai mare decât gradele de libertate, poate fi un indiciu al supradispersiei.

Se poate rezolva o posibilă supradispersie prin aplicarea procedurii de cvasi-probabilitate, care permite estimarea parametrilor modelului fără o cunoaștere completă a distribuției erorii variabilei de răspuns. Procedura poate permite calcularea lui *j* din date (mai degrabă decât să presupunem că este unitate pentru distribuțiile Poisson și binomiale).

Pentru a aplica estimarea cvasi-probabilității la polipii de date colonici3 , pot fi utilizate următoarele segmente de cod R :

```
> install.packages(„HSAUR")
> bibliotecă (HSAUR)
> ls("pachet:HSAUR")
> polipi3
> polyps3_GLM_Quasi <- glm(number3m ~ tratament + varsta,
+                          date = polipi3, familie = quasipoisson() )
> rezumat(polyps3_GLM_Quasi)
```

Apel:
glm(formula = număr3m ~ tratament + vârstă, familie = quasipoisson(), date = polipi3)
Reziduuri de devianta:

Min	1Q	Median	3Q	Max
-11,3117	-5,0824	-2,7838	-0,4611	22,4333

Coeficienți:

	Estima	Std. Eroare	valoare	Pr(>|t|)
(Intercepta)	5,72284	1,16148	4.927	9.36e-05 ***
Tratamentactiv	-1,19595	0,74361	-1.608	0,124
Vârstă	-0,07243	0,05145	-1.408	0,175

Signif. coduri: 0 '*' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 " 1**

(Parametrul de dispersie pentru familia quasipoisson considerat a fi 91,14169)

Devianța nulă: 1603,1 pe 21 de grade de libertate
Devianța reziduală: 1191,6 pe 19 grade de libertate AIC: NA

Numărul de iterații Fisher Scoring: 6
>

(a) Comparați coeficienții de regresie pentru ambele variabile explicative: tratament și vârstă .
   (i) Care sunt valorile lor?
   (ii) Este fiecare valoare semnificativă?

(b) Comparați erorile standard estimate pentru ambele variabile explicative:
   tratament (0,74316) și vârstă (0,05145)
   cu valorile calculate anterior pentru distribuția Poisson simplă: tratament (0,077891) și vârstă (0,005389).

(c) Este evident că dispersiile pentru cazul cvasi-probabilității sunt mult mai mari. De ce?

(d) O explicație sugerată pentru o astfel de supradispersie a datelor este că polipii colonici pot apărea în grupuri (adică, ei nu apar independent). Comentează această sugestie.

4. ***Funcțiile*** predict() ***și*** predict.glm() .

Un instrument biostatistic util, bazat pe regresia logistică, este funcția predict() , care are mai multe variații, inclusiv funcția predict.glm() . Acestea sunt documentate în statisticile pachetului CRAN .

Fiecare dintre aceste funcții R poate fi utilizată pentru a calcula predicții și pentru a estima erori standard (printre altele) ale predicțiilor dintr-un obiect model liniar generalizat adaptat.

Utilizare:

```
prezice(obiect, date noi = NULL,
        tip = c(„link", „răspuns", „termeni"),
        se.fit = FALSE, dispersie = NULL, termeni = NULL, na.action = na.pass, ...)
```

Argumente:

obiect	Un obiect adaptat al clasei care moștenește de la „glm" .
newdata	Opțional, un cadru de date în care să căutați variabile cu care să preziceți.
tip	Tipul de predicție necesar. Valoarea implicită este pe scara predictorilor liniari; „răspunsul" alternativ este pe scara variabilei răspuns.
se.fit	Un comutator logic care indică dacă sunt necesare erori standard.
dispersie	Se presupune că dispersia modelului liniar generalizat se potrivește în calcularea erorilor standard. Dacă este omis,
termeni	Cu type="terms" , în mod implicit toți termenii sunt returnați. Un personaj vector specifică ce termeni vor fi returnați.
na.action	Funcția care determină ce ar trebui făcut cu valorile lipsă din newdata . Implicit este de a prezice NA.
...	Argumente suplimentare transmise către sau de la alte metode.

Valoare:

Dacă se = FALSE , un vector sau o matrice de predicții.

Dacă se = TRUE , o listă cu componente.

Fit Predictii.
se.fit Erori standard estimate.
rezidual.scale Un scalar care dă rădăcina pătrată a dispersiei utilizate în calculul erorilor standard.
glm, SafePrediction

Exemple:

necesită (grafică)

```
>   # exemplu din Venables și Ripley (2002, pp. 190–192.)
ldoză <- rep(0:5, 2)
numdead <- c(1, 4, 9, 13, 18, 20, 0, 2, 6, 10, 12, 16)
sex <- factor(rep(c("M", "F")), c(6, 6)))
SF <- cbind(numdead, numalive=20-numdead)
budworm.lg <- glm(SF ~ sex*ldoză, familie=binom)
rezumat(budworm.lg)

plot(c(1,32), c(0,1), tip = „n", xlab = „doză",
        ylab = "prob", log = "x")
text(2^ldoză, numdead/20, ca.caracter(sex))
ld <- seq(0, 5, 0,1)
lines(2^ld, predict(budworm.lg, data.frame(ldose=ld, sex=factor(rep(„M",
        lungime(ld))), levels=levels(sex))), tip = „răspuns" ))
lines(2^ld, predict(budworm.lg, data.frame(ldose=ld, sex=factor(rep(„F",
        length(ld)), levels=levels(sex))), tip = „răspuns" ))
```

Următorul segment de cod R preia execuția acestui exercițiu exemplu:

```
>
>   necesită (grafică)
>
>   ## exemplu din Venables și Ripley (2002, pp. 190-192.)
>   ldoză <- rep(0:5, 2)
>   numdead <- c(1,4, 9, 13, 18, 20, 0, 2, 6, 10, 12, 16)
>   sex <- factor(rep(c("M", "F")), c(6, 6)))
>   SF <- cbind(numdead, numalive=20-numdead)
> budworm.lg <- glm(SF ~ sex*ldose, family=binoal)
>   rezumat(budworm.lg)
```

Apel:
glm(formula = SF ~ sex * ldoza, familie = binom) Reziduuri de devianta:

Min	1Q	Median	3Q	Max
-1,39849	**-0,32094**	**-0,07592**	**0,38220**	**1,10375**

Coeficienți:

	Estima	Std. Eroare	valoarea	Pr(>|z|)
(Intercepta)	**-2,9935**	**0,5527**	**-5.416**	**6.09e-08 *****
sexM	**0,1750**	**0,7783**	**0,225**	**0,822**
ldoza	**0,9060**	**0,1671**	**5.422**	**5.89e-08 *****

sexM:ldoză 0,3529 0,2700 1.307 0,191

Signif. coduri: 0 '*' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 " 1**

(Parametrul de dispersie pentru familia binomială considerat a fi 1)

Devianța nulă: 124,8756 pe 11 grade de libertate
Devianța reziduală: 4,9937 pe 8 grade de libertate

AIC: 43.104
Numărul de iterații Fisher Scoring: 4

```
>
>   # Pregătirea graficului și a etichetelor pe cele două axe:
>   plot(c(1,32), c(0,1), tip = „n", xlab = „doză",
+                                    ylab = "prob", log = "x")
>   text(2^ldoză, numdead/20, ca.caracter(sex))
>   ld <- seq(0, 5, 0,1)
>
>   # Desenarea liniei predict() pentru bărbați:
> linii(2^ld, predict(budworm.lg, data.frame(ldose=ld,
+ sex=factor(rep("M", lungime(ld)), niveluri=nivele(sex))),
+ tip = „răspuns"))
>
>   # Desenarea liniei predict() pentru femei:
>   lines(2^ld, predict(budworm.lg, data.frame(ldose=ld,
+ sex=factor(rep(„F", lungime(ld)), niveluri=nivele(sex))),
+ tip = „răspuns"))
>
>   # ieșire : Figura 7.33 prezice()
>
```

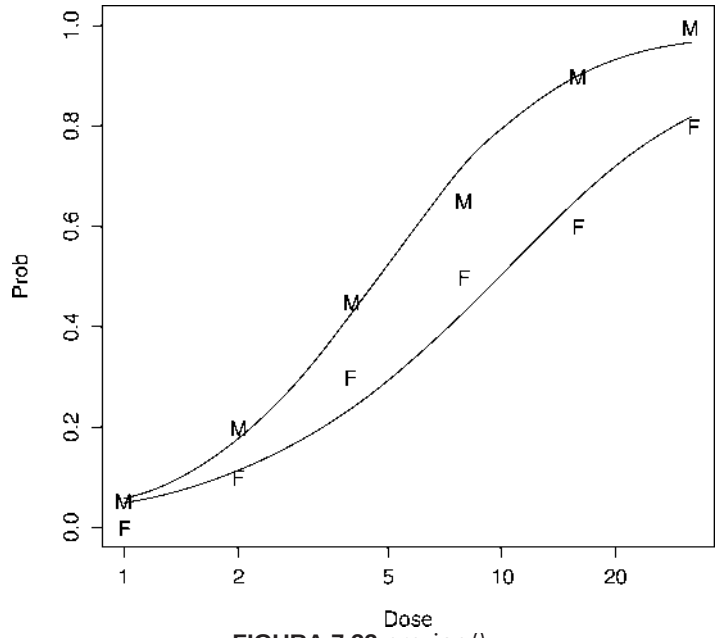


FIGURA 7.33 prezice().

(a) În acest program, când se utilizează funcția predict() , în ce stadiu este introdusă funcția de regresie liniară glm() ?
(b) Care este funcția de legătură folosită în model? Este aceasta o alegere bună? De ce sau de ce nu?
(c) Cum și când a fost introdusă funcția de legătură în calculul modelului?
(d) Comentați mărimile și semnificațiile erorilor standard ale variabilelor dependente pentru regresie.
(e) Propuneți o abordare eficientă pentru a constata aceste erori.

5. *Regresia logistică în analiza supraviețuirii: date despre infecția cu boala granulomatoasă cronică* (cgd) .

În analiza supraviețuirii, procedura K–M este utilizată pentru a estima probabilitatea de supraviețuire pentru o anumită perioadă de timp. Calculele constau în calcularea proporțiilor subiecților de caz dintr-un eșantion care supraviețuiesc pentru diferite durate de timp. Aceste proporții ale eșantionului sunt apoi utilizate ca estimări ale probabilităților de supraviețuire pe care s-ar aștepta să se observe în tipul de populație reprezentat de eșantionul de testare. În statistica matematică, această abordare este cunoscută ca o *tehnică neparametrică,* spre deosebire de funcțiile de legătură comune (în care sunt presupuse Poisson, binom sau alte probabilități) utilizate într-o regresie logistică de bază.

Setul de date cgd din Fleming și Harrington (1991) provine dintr-un RT controlat cu placebo al interferonului gamma în boala granulomatoasă cronică (CGD). Un total de 128 de subiecți din 13 spitale au fost urmăriți timp de aproximativ un an. Numărul de subiecți cazați per spital a variat de la 4 la 26. Fiecare dintre ele ar fi putut suferi mai mult de o infecție. Timpii de supraviețuire (timpi până la eveniment) sunt timpii dintre infecțiile recurente cu CGD la fiecare pacient (adică, timpii interval). Cenzurarea a avut loc la ultima observație pentru toți pacienții cu excepția unuia, care a suferit o infecție gravă la data la care subiectul cazului a părăsit studiul.

Utilizare:
date (cgd)

Format:
Setul de date CGD conține 15 coloane și 203 rânduri. O scurtă descriere a coloanelor de date este următoarea:

Id	Număr de pacient pentru 128 de subiecți de caz.
Centru	Număr centru de înscriere pentru 13 spitale.
Aleatoriu	Data randomizării.
Trata	Tratament cu interferon gamma (rIFN-g) sau placebo (Placebo).
Sex	Sexul fiecărui subiect de caz (bărbat, femeie).
Vârstă	Vârsta fiecărui subiect de caz la intrarea în studiu, în ani.
Înălțime	Înălțimea fiecărui subiect de caz la intrarea în studiu, în centimetri.
Greutate	Greutatea fiecărui subiect de caz la intrarea în studiu, în kilograme.
Moşteni	Model de moștenire (autosomal recesiv, legat de X).
Steroizi	Utilizarea corticosteroizilor la momentul intrării în studiu (1 = Da, 0 = Nu).

Proylac	Utilizarea antibioticelor profilactice la momentul intrării în studiu (1 = Da, 0 = Nu).
hos.cat	O clasificare a regiunii spitalicești în patru grupuri.
Tstart	Începutul fiecărui interval de timp.
Enum	Numărul de ordine. Pentru fiecare pacient, înregistrările de infecție sunt în ordinea numărului de ordine.
Oprește-te	Sfârșitul fiecărui interval de timp.
Stare	Indicator de cenzură (1 = necenzurat, 0 = cenzurat).

Referinte:
Fleming și Harrington (1991).
Theneau și Grambsch (2001) .

Următorul segment de cod R preia execuția acestui exercițiu exemplu:

```
>
> install.packages(„packHV")
> bibliotecă (packHV)
```

Se încarcă pachetul necesar: WriteXLS

```
> ls("pachet:pachetHV")
```

[1] „compara" „convert_factor" „convert_zero_NA" „desc"
[5] „hist_boxplot" „IC_OR_glm" „IC_RR_coxph" c „multitable"
[9] "plot_km" c "plot_mm" "plot_multi.table" ccc"plot_reg"

```
> IC_OR_glm
```

(Ieșire omisă)

```
> cgd
```

	centru de identificare	aleatoriu	trata	sex	vârs	înălțime	greutat
1	1 Institutul Scripps	1989-06-07	rIFN-g	femeie	12	147,0	62,0
2	1 Institutul Scripps	1989-06-07	rIFN-g	femeie	12	147,0	62,0
3	1 Institutul Scripps	1989-06-07	rIFN-g	femeie	12	147,0	62,0
203	135 Institutul Scripps	29.12.1989,	femeie	placebo	3	96,0	13.1

	moștenește	steroizi	propilac	hos.cat	tstart	enumerare	tstop	starea
1	autozomal	0	0	SUA:	0	1	219	1
2	autozomal	0	0	SUA:	219	2	373	1
3	autozomal	0	0	SUA:	373	3	414	0
203	autozomal	0	1	SUA: altele		1	227	0

```
>
> cgd$time=cgd$tstop-cgd$tstart
> plot_km(Surv(time,status)~sex,data=cgd,lwd=2,
+ col=c("negru","rosu"))
> # Ieșire : Figura 7.34 .
```

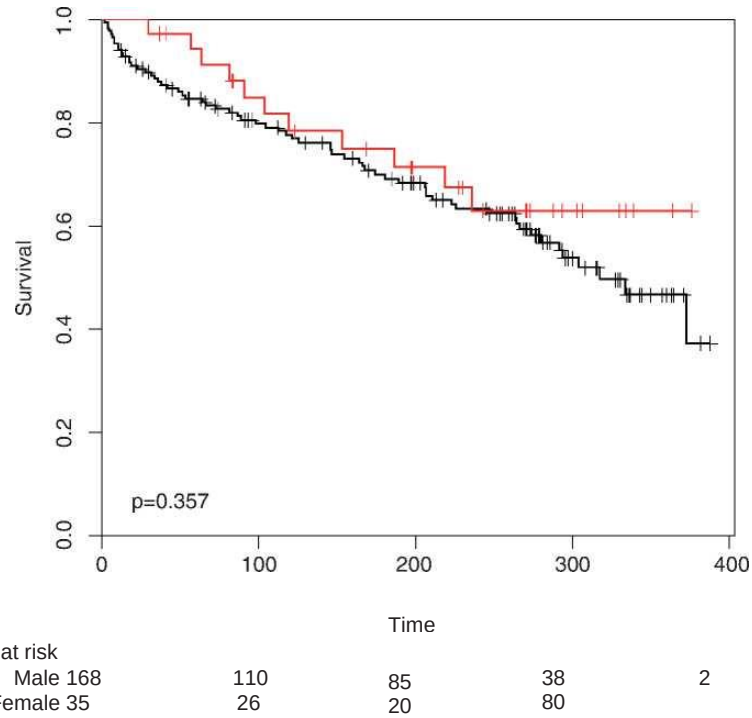


FIGURA 7.34 Regresia logistică: analiza de supraviețuire a datelor cgd . Parcele: masculin (curba inferioară) și feminin (curba superioară).

(a) În analiza de supraviețuire, ce funcție de legătură este utilizată în regresia logistică rezultată?
(b) Privind figura 7.34 , ce se poate concluziona cu privire la funcțiile de legătură pentru cele două populații: subiecții caz masculin și feminin?
(c) Dacă „pașii" curbelor de supraviețuire ar fi neteziți, la ce tip de funcții de legătură ar putea fi de așteptat?
(d) În Figura 7.34 , curba roșie (pentru subiecții de sex feminin) se află *deasupra* curbei negre (pentru subiecții de sex masculin). Cum se reflectă această caracteristică asupra probabilităților de supraviețuire ale acestor două cohorte?
(e) Deoarece ambele curbe arată o pantă descrescătoare în raport cu timpul, ce familie de curbe de distribuție a probabilității poate fi utilizată pentru a modela procesul de supraviețuire?

REFERINȚE

Blanchard, MS (2012). Pachetul CRAN PKtools . sblanchard@coh.org

Calvert, AH și Plummer, R. (2008, 15 iunie). Dezvoltarea metodologiilor de fază I pentru studiile de cancer : utilizarea punctelor finale farmacocinetice și farmacodinamice creează scena pentru studiile clinice de fază 0 pentru cancer. *Clinical Cancer Research, 14* , 3664.

Spitalul Milostivirea Copiilor. (2000). *Ce este coliniaritatea?* Preluat din „Ask Professor Mean", Categorie: Probleme de modelare; http://www.childrensmercy.org/stats/ask/collinearity.aspx

Dalgaard, P. (2002). *Statistici introductive cu R* (Springer Series in Statistics and Computing). New York, NY: Springer.

Daniel, WW (2005). *Biostatistică: O fundație pentru analiză în științele sănătății* (ed. a 7-a; Seria Wiley în Probabilitate și Statistică). New York, NY: Wiley.

Everitt, BS și Hothorn, T. (2006). *Un manual de analize statistice folosind R* (HSAUR). Boca Raton, FL: Chapman & Hall/CRC.

Fitzmaurice, G. (2004). Ajustarea pentru confuzie. *Nutrition, 20* (6), 594596. Preluat de la http://epicentre.massey.ac.nz/Portals/0/EpiCentre/Downloads/Education/227-407/ Fitzmaurice_2004.pdf/

Fleming, TR și Harrington, DR (1991). *Procese de numărare și analiză de supraviețuire* . New York, NY: Wiley.

Gerds, TA, Cai, T. și Schumacher, M. (2008). Performanța modelelor de predicție a riscului. *Biometrical Journal, 4* , 457479.

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Evaluarea și compararea schemelor de clasificare prognostică pentru datele de supraviețuire. *Statistics in Medicine, 18* , 25292545.

Hastie, TJ și Tibshirani, RJ (1990). *Modele aditive generalizate.* Boca Raton, FL: Chapman & Hall/CRC.

Heagerty, PJ, Lumley, T. și Pepe, MS (2000). Curbe ROC dependente de timp pentru date de supraviețuire cenzurate și un marker de diagnostic. *Biometrie, 56* , 337344.

Katki, HA și Mark, SD (2008). Analiza supraviețuirii pentru cohortele cu informații despre covariate lipsă. *R-News* , *8* (1), 14–19. Preluat de la http://dceg.cancer.gov/Files/RNewsNestedCohortWrapper_old.pdf

Katki, HA și Mark, SD (2009). Analiza de supraviețuire a studiilor imbricate în cohorte folosind pachetul NestedCohort . CRAN (The Comprehensive R Archive Network). https://CRAN.r- project.org/

McNamee, R. (2005). Modelarea regresiei și alte metode de control al confuziei. *Occupational & Environmental Medicine, 62* , 500506. Preluat de la http://oem.bmj.com/continut/62/7/500.full

Moreira, A., Araujo, AA și Machado, LM (2012). Supraviețuire pachet CRANBIV . Preluat de pe http://cran.r-project.org/

Schnute, JT, Couture-Beil, A., Haigh, R. și Kronlund, AR (2011). *PBSmodelling 2.62: Ghidul utilizatorului* . Nanaimo, British Columbia, Canada: Fisheries and Oceans Canada, Science Branch. CRAN (The Comprehensive R Archive Network). https://CRAN.r-project.org/

Stanley, K. (2007). Introducere statistică pentru cercetarea cardiovasculară: Proiectarea trialurilor randomizate controlate. *Tiraj, 115* , 1164–1169.

Sigler, C. și Stemhagen, A. (2011). Epidemiologia în dezvoltarea medicamentelor. Preluat de pe http://www.pharmasug.org/proceedings/2011/HS/PharmaSUG-2011-HS10.pdf

Therneau, TM și Grambsch, PM (2001). Modelarea datelor de supraviețuire: Extinderea modelului Cox. *Statistici pentru Biologie și Sănătate* . New York, NY: Springer.

Tornoe, CW (2012). Pachetul CRAN nlmeODE . ctornoe@gmail.com

Venables, WN și Ripley, BD (2002). *Statistică aplicată modernă cu S* (ed. a IV-a, pp. 190–192). New York, NY: Springer Science & Business Media.

Wood, SN (2006). *Modele de aditivi generalizate: O introducere cu R.* Boca Raton, FL: Chapman & Hall/CRC.

Index

randomizare adaptativă a monedelor părtinitoare, 347
Studiul Adventist de Sănătate 1 (AHS-1), 13
Studiul Adventist de Sănătate 2 (AHS-2), 13
ascunderea alocării, 346
Test de screening pentru diagnosticul bolii Alzheimer, 230–231
Analiza variației (ANOVA)
  epidemiologie nutrițională, 148–150
  populația înseamnă 141
  reziduuri, 141
  regresie liniară simplă, 141
epidemiologie analitică, 8–9
Medicament antiepileptic (AED) Sarcina Registru, 364
tablouri și matrici, 108–109
Argumentul aspectului, complot latice, 189
sisteme automate de codificare, 83
topoare, 166–167

Vaccinul Bacil Calmette-Guerin (BCG). eficacitate, 325–327
scanere de coduri de bare, 84
grafică de bază (sau tradițională), 155–156
  drivere de dispozitiv, 168–170
  elemente, 166
  marjele figurii, 167
  funcții de trasare la nivel înalt
    argumente ale, 160–161
    funcția plot() , 157
    seturi de date multivariate, 157–160
  funcția de identificare, 165
  locator, 165
  funcții de trasare la nivel scăzut
    abline și ax, 161
    legenda, 162–163
    expresii matematice, 164
    puncte și linii, 161
    text, 161
    titlu, 162
  ieșiri cu mai multe cifre, 168
  parametri, 165–166
  funcția plot() , 166
  marcaje și topoare, 166–167
codificare lot, 84
Biostatistica bayesiană, 268
Interval Bayesian credibil, 48–49
Inferența bayesiană, 40, 269–270
teorema lui Bayes
  Inferența bayesiană, 269–270
  Modelul de regresie Cox, 284–288
    funcția de pericol de bază, 285
    funcția de pericol, 284
    regresia hazardelor proporționale, 285
    R , 286–288
    analiza supraviețuirii, 285–286
  funcția de pericol cumulativ, 269
  screening diagnostic, 229–231
  formula Bayes inversă (*vezi* Bayes invers formula (IBF))
  Procedura Kaplan-Meier
    Fișier de date privind leucemia mielogenă acută, 281–283
    subiecte de caz, 273
    radioimunoterapie gliom, 274–281
    funcția survfit() , 273
    Funcția Surv() , 273–274
    funcția de supraviețuire, 272–273
  model de regresie liniară multiplă, 283–284
  în teoria probabilității, 268
calculator de distribuție binomială, 59–60
model binom, 21–22
probabilitate binomială, 58
risc binom, 68–69
Bioconductor, 370
biostatistica (BIOS)
  ANOVA, 141, 148–150
  aplicații ale, 50–51
  model binom, 21–22
  risc binom, 68–69
  Pachetul CRAN, 299–304
  analiza datelor
    date privind rezultatele continue, 348
    date de rezultat dihotomice (binare), 348
    determinarea dimensiunii eșantionului, 349–350
    surse, 348
    date despre rezultatul timpului până la eveniment, 348

biostatistică (BIOS) (*cont.*)
demografice, 336, 338, 339
descriptiv, 22
prevalența bolii, 21
analiza datelor epidemiologice
măsuri de frecvență, 52
evenimente legate de sănătate, 56
estimare punctuală, 56–57
raport, 52–55
Testul GoF (*vezi* testul de bunăstare a potrivirii)
testarea ipotezelor (*vezi* testarea ipotezelor și valorile p)
deducere
Interval Bayesian credibil, 48–49
interval de încredere (*vezi* intervalul de încredere (IC))
BIOS frecventist, 41–42
măsurători de interval, 41
moduri de, 40
măsurători nominale, 41
măsurători ordinale, 41
măsurători de raport, 40
inferențial, 22
model neparametric, 298
model parametric, 298
Date privind numărul și rata de Poisson, 68
date de prevalență, 69
probabilitate (*vezi* probabilitate)
Mediul R
activități de calcul, 89–90
Versiunea R-2.14.1, 87
Mediu versiune R-2.9.1, 87
Limbi S/S-PLUS, 87
percentilele eșantionului, 232
model semi-neparametric, 299
model semiparametric, 298
niveluri de semnificație, 237–240
măsură unică de apariție, 67–68
rezumat, 232–233
două măsuri de apariție, 69–70
analiza datelor bivariate
distribuție normală, 138
Pachete R , 136
calculul mărimii eșantionului, 137
distribuție normală bivariată, 138
studii randomizate orb, 347
indicele de masă corporală (IMC), 102–103
criteriile Bradford-Hill
coerență, 28
consistenta, 27
relația doză-răspuns, 27 experiment și
specificitate, 27–28 plauzibilitate, 27
Comentariile lui Rothman, 28
puterea de asociere, 27
relație temporală, 26

studiu caz-control, 14, 376–377
avantaje, 324
Eficacitatea vaccinării BCG, 325–327
limitări, 325
studiu observațional, 323
teorema limitei centrale (CLT), 61–62
prejudecată de canalizare, 364
distribuția chi-pătrat, testul GoF
distribuții continue
distribuție empirică, 248
Testul K–M GoF, 250–252
R , 249–250
Testul Shapiro–Wilk (SW) GoF, 252–253
omogenitate, 247–248
independență pentru GOF, 244–247
frecvența observată și așteptată, 241
R , 242–244
varianța eșantionului, 241
epidemii de holeră, 10–12
probabilitate clasică, 212
modelul clasic de regresie, 136
studii clinice, 329
studiu de cohortă, 377
avantaje, 329
Studii de sănătate adventiste, 13
cerere, 329
cohorta de naștere, 328
grup de comparație, 328
modelul de regresie proporțională Cox, 332–336
boală și expunere, 329
Studiul de sănătate Framingham, 13
Diagrama Lexis, 336, 338
studiu longitudinal, 323
boli netransmisibile, 330–332
Studiul de sănătate a asistentelor medicale, 13
retrospectivă, 13
participanții la studiu, 12
Pachetul complet R Archive Network (CRAN).
coxphf, set de date despre cancerul de sân, 328
coxrobust, 332
CRTSize, 350–352
gam AND gamair, 300–304
grofit, curbele doză-răspuns, 299
NestedCohort, 393–399
nlmeODE, farmacocinetica indometacinei, 371–375
PKtools, 370–372

randomSurvivalForest, 352–362
dimensiunea eșantionului, 349–350
supraviețuireBIV, 399–412
supraviețuire, fișier setul de date privind vezica urinară, 328
supraviețuireROC, 389–393
survrec, 341–343
estimare condițională de maximă probabilitate (CMLE), 55
probabilitate condiționată, 216–217
interval de încredere (IC)
criterii de acceptare, 237
Rata de deces direct ajustată în funcție de vârstă, 46–47
populații finite, 45–46
rate ajustate indirect în funcție de vârstă, 48
teste t cu una și două probe, 236–237
Distribuția Poisson, 46
interpretare practică, 234
interpretare probabilistică, 233–234
Loturi Q–Q, 235–236
cuantile și mediană, 234–235
evenimente rare, 46
rata este egală cu zero, 46
distribuția prin eșantionare a mediei, 43
eșantioane mici, 44–45
eroare standard, 44
teoria eșantionării statistice, 42
distribuția T a elevului, 45
scor z, 43
confuzător, 375–376
confuză, analiză de supraviețuire
ajustare
modelarea regresiei (*vezi* regresia modelare, ajustare de confuzie)
stratificare (*vezi* stratificare, ajustare confuză)
studii caz-control, 376–377
studii de cohortă, 377
coliniaritate, 386–388
confuzător, 375–376
dublu orb, 377
studiu randomizat controlat, 377
vedere schematică, 376
stratificare, 377–378
model contrastiv
cauze, 16
H. pylori în formarea ulcerului, 16–17
simptome, 15
model de corelare, 136
Modelul de regresie Cox
funcție de pericol de bază, 285 funcție de pericol, 284
date despre cancerul pulmonar, 232–338
regresia hazardelor proporționale, 285
Segment de cod R , 286–288 analiză de supraviețuire, 285–286 set de date despre fumul de tutun, 331 valoare critică, 63–64 studii transversale, 14

baze de date, 87
codificarea datelor, 83
introducerea datelor și analiza funcției
data.entry() , 119 cadre de date
consta din, 116
constructii, 129
creație, 116
editare, 125–127
foaie de calcul Excel, 116–118
FUNCTION list() , 127–129 fișier text, 118–119, 130
analiza multivariată a datelor, 138–140
R , 120–121
Programarea R (*vezi programarea* R)
funcțiile scan() , 122–123 funcții sursă () , interfață de foi de calcul 123–124 ,
analiză univariată a datelor 124–125
segmente de cod, 133
funcția reexp() , 133, 134
R , 134–135
Funcția univax() , 134
prelucrarea datelor
captura de date, 84
codificarea datelor, 83 editarea, 84–85
imputări, 85 producerea de rezultate, 87
controlul calității, 85–87
biostatistica demografică, 336, 338, 339
biostatistică descriptivă, 22
epidemiologie descriptivă, 8 funcția dev() , 169
drivere de dispozitiv, 168–170
Studiul pentru controlul și complicațiile diabetului (DCCT), 18
Rata de deces direct ajustată în funcție de vârstă (DAADR), 46–47
cauzarea bolii
criteriile Bradford-Hill
coerență, 28
consistenta, 27
relația doză-răspuns, 27
experiment și specificitate, 27–28
plauzibilitate, 27
puterea de asociere, 27
relație temporală, 26
criterii cauzale, 38
inferențe cauzale, 37–38
boli cronice netransmisibile, 14

cauze componente și plăcinte cauzale, 30–31
model contrastiv, 15–17
mediu, 30
triunghi epidemiologic, 30
agent extern, 29–30
cauze genetice și de mediu, 35–37
interacțiunea între cauze, 32
model monocauzal, 15
model multifactorial, 15
Diagramele lui Rothman, 33–35
dovezi științifice, 38–39
puterea unei cauze, 31–32
cauză suficientă, 33
gazdă susceptibilă, 30
boală (morbiditate) odds ratio (DOR), 70–71
dublu orb, 377

modificarea efectului, 384
funcția de supraviețuire empirică (esf), 282–283
Epidata, 130
epidemiologie (EPDM)
analitic, 8–9
cerere, 7
risc binom, 68–69
bios, 21–22
biostatistică (*vezi* biostatistică)
povestea cu lebedele negre, 17–18
Criteriile Bradford-Hill, 26–28
studiu caz-control (*vezi* studiu caz-control)
studiu de cohortă (*vezi* studiul de cohortă)
definiție, 2, 6
descriptiv, 8
determinanti, 6
cauzalitatea bolii (*vezi* cauzalitatea bolii)
distribuție, 6
studiu experimental, 10
măsurarea expunerii în, 19–20
măsuri de frecvență, 52
stări/evenimente legate de sănătate, 6–7
interpretare juridică, 28–29
natura *vs.* hrăni, 18–19
studiu observațional, 10–14
Date privind numărul și rata de Poisson, 68
date de prevalență, 69
raportul ratei, 70
raport, 52–55
raportul de risc și DOR, 70–71
Studiul epidemiologic al vaccinului antipolio Salk, 72
cercetare științifică, 6
măsură unică de apariție, 67–68
populații specificate, 7
analiza supraviețuirii (*vezi* analiza supraviețuirii)
sarcini ale, 9
două măsuri de apariție, 69–70
test binom exact, 60

fluorizarea apei potabile, 2–3
epidemiologie criminalistică, 29
Studiul de sănătate Framingham, 13
biostatistica frecventistă
semnificație biostatistică, 42
estimarea intervalului, 41

model aditiv generalizat (GAM), 300–304
grafic X–Y generic
argumente, 152
detalii, 152–153
utilizare, 152
testul de bunăstare a potrivirii (GoF).
distribuția chi-pătrat
distribuții continue, 248–253
omogenitate, 247–248
independență pentru GOF, 244–247
frecvența observată și așteptată, 241
R , 242–244
varianța eșantionului, 241
investigații epidemiologice, 241
fitdistr() , 254–257
distribuții normale, 241
testarea ipotezelor statistice, 240
grafică grilă, 156
epidemiologia cancerului, 203
gazele de evacuare ale motorului, 194–197
epidemiologia mediului, 197–202
primitive grafice, 191–192
grafică lattice, 183–189
obiecte, 192–194
R grafică, 189–192

cercetare și politică în domeniul sănătății (HRP), 2–3
cercetarea serviciilor de sănătate (HSR), 3
inaltime, 169
HSAUR, 130
hiperplan, 139
Limbajul de marcare hipertext (HTML)
interfață, 90
testarea ipotezelor și valorile p
calculator de distribuție binomială, 59–60
probabilitate binomială, 58
teorema limitei centrale, 61–62
valoare critică, 63–64
calcul direct, 59
test binom exact, 60

eșantion mare, 62–63
marja de eroare, 65
ipoteza nulă, 58
putere, 60
R , 67
dimensiunea eșantionului necesară, 66–67

imputații, 85
rata de deces sau îmbolnăvire standardizată indirect (ISR), 48
biostatistică inferențială, 22
captura interactivă a datelor, 84
formula Bayes inversă (IBF)
presupunere continuă, 290
identitate pdf comună, 289, 290
sprijin comun, 289, 290
MCMCregress() , 293–298
spațiu măsurabil neprodus, 288
punctual IBF, 292–293
funcția de densitate de probabilitate, 289
spațiu măsurabil al produsului, 288, 291

probabilitate comună, 217

Procedura Kaplan–Meier (K–M).
subiecte de caz, 273
regresie logistică, 423
funcția survfit() , 273
Funcția Surv() , 273–274
analiza supraviețuirii
Fișier de date privind leucemia mielogenă acută, 281–283
supraviețuirea cancerului, 399
radioimunoterapie cu gliom, 274–281
funcția de supraviețuire, 272–273
set de date despre fumul de tutun, 332

diagramă grafică grilă grilă
aranjamente, 188–189
epidemiologia cancerului, 203
controlând, 186–188
Diagrame și suprafețe de dispersie 3D, 185–186
funcția xyplot() , 184–185
argument de aspect, diagramă lattice, 188
Diagrama Lexis, 336, 338
regresii logistice
analiză, 414–415
variabilă dependentă categorică binară, 412
date de infecție cu boala granulomatoasă cronică, 434–437
modele liniare generalizate, 415–419
transformare logit, 414
model, 413–414
modele, 386
variabilă dependentă categorică multinomială, 412
predict() și predict.glm() , 431–434
validarea probabilității prezise, 423–430
curba de supraviețuire, 420–423
sistem univariat, 413
Starea lui Lyapunov, 62 de ani

macroeditare, 85
înregistrări magnetice, 84
probabilitate marginală, 222–223
marja de eroare, 65
statistici matematice
interval de încredere (*vezi* intervalul de încredere (IC))
Testul GoF (*vezi* testul de bunăstare a potrivirii)
percentilele eșantionului, 232
niveluri de semnificație, 237–240
rezumat, 232–233
procesul expirator maxim (MEP), 145
mediană, 234–235
estimarea mediană imparțială (MUE), 55
Studiul Medicare privind poluarea aerului (MCAPS), 150
medicina, 1
microeditare, 85
model monocauzal, 15
model multifactorial, 15
model de regresie liniară multiplă, 283–284
analiza modelului de regresie multiplă, 140
analiza multivariata a datelor
analiza modelului de corelație, 139–140
Studiu epidemiologic de fibroză chistică, 142–147
analiza regresiei liniare, 138–139
epidemiologie nutrițională, 148–150

imputarea celui mai apropiat vecin, 85
ipoteza nulă, 58, 245–246

probabilitate obiectivă, 212
studii observaționale EPDM
studiu caz-control, 14
epidemii de holeră, 10–12
studii de cohortă, 12–14
studii transversale, 14
epidemiologie observațională a HP
stilul de viață, 3
test t unilateral, 237

cititoare optice de caractere, 84

funcția par() , 165–166
grafica par() , 166
PBSmodelling, 370
randomizare cu blocuri permutate, 347
sănătatea personală, 3–4
farmacoepidemiologie, 363–364
dezvoltarea fazei
  dezvoltarea medicamentelor
    faza I, 363
    faza II, 363
    faza III, 363
    faza 0/faza preclinica, 361
  epidemiologic, 364–369
  farmacoepidemiologie, 363–364
funcția plot() , 166
puncte() , 166
Distribuția Poisson, 46
Modelul de regresie Poisson, 386
probabilitate posterioară, 270
diagrame postscriptive, 169
p-quantila, 234
medicina preventivă, 1–2
probabilitate
  regula adăugării, 218–220
  alege() , 224
  clasic, 212
  evenimente complementare, 215–216
  condițional, 216–217
  test de screening diagnostic
    Teorema lui Bayes, 229–231
    probabilitate condiționată, 228
    starea bolii, 227–228
    valoare predictivă negativă, 228–229
    valoare predictivă pozitivă, 228
    specificitate și sensibilitate, 228
  proces de testare clinică dihotomică, 227
  funcția factorial() , 223
  independența și dependența de apariții, 220–222
  comun, 217
  marginal, 222–223
  regula înmulțirii, 218
  evenimente care se exclud reciproc, 213
  obiectiv, 212
  funcția prod() , 225–226
  proprietatea exhaustivității, 213
  frecvență relativă, 212
  sample() , 224–225
  spațiu de probă, 213–215
  subiectiv, 212
funcția de densitate de probabilitate (pdf), 289

studii prospective de cohortă, 13
studiu de cohortă prospectiv, 364
sănătate publică (PH)
  definiție, 2
  epidemiologie observațională, 3
  vs. sănătatea personală, 4–5
  politica, 4
  recomandari, 4

asigurarea calității, 86
controlul calității, 86
cuantile, 234–235

randomizare
  avantaje, 346
  proceduri, 346
  restrâns, 347
  simplu, 347
studii randomizate controlate (RCT), 329–330, 377
studii randomizate (RT)
  analiza datelor biostatistice
    date privind rezultatele continue, 348
    date de rezultat dihotomice (binare), 348
    determinarea mărimii eșantionului, 349–350 surse, 348
    date despre rezultatul timpului până la eveniment, 348
  orbit, 347
  cluster, 345
  crossover, 345
  factorial, 345–346
  grup paralel, 345
  randomizare, 346–347
  neorbit, 347–348
raportul ratei, 70
raportul de risc reciproc, 72
analiza de regresie, 136
modelare de regresie, ajustare de confuzie
  fumători de țigări, 384–385
  comparație de grup, 385–386
  modele de regresie logistică, 386
  subiecte de caz nefumători
    ieșire computer, 383
    VEMS și vârstă, 382, 383
    relație liniară, 383
    metode statistice, 383
  Modelul de regresie Poisson, 386
  pașii implicați în, 385
probabilitate relativă de frecvență, 212
Mediul R
  biostatistica
    activități de calcul, 89–90

Versiunea R-2.14.1, 87
Mediu versiune R-2.9.1, 87
Limbi S/S-PLUS, 87
calcule, 100–101
constă din, 87–88
caracteristici ale, 90–98
operații matematice, 99–100
randomizare restrânsă, 347
studiu de cohortă retrospectiv, 13–14, 364
grafica R
grafică de bază (*vezi* grafica de bază (sau tradițională))
alegerea, 156
grafică grilă (*vezi* grafică grilă)
niveluri, 155
pachete, 156–157
diferență de risc (DR), 69
raportul de risc (RR), 69
programare R
indicele de masă corporală (IMC), 102–103
variabile categoriale, 101
matrici
matrice și, 108–109
colnames, rownames și t, 110–111
funcția de dimensiune, 109
matricea funcției, 109
R , 102
cuvinte rezervate, 111
grafică simplă, 103–106
vectori
vector de caractere, 107
funcția „concatena”, 108, 111
vector logic, 107–108
valori lipsă, 111
Funcția „ rep ”, 108, 112
Funcția „ seq ”, 108, 111–112

Studiu epidemiologic al vaccinului antipolio Salk, 72 eroare de prelevare, 42 funcțiile scan() , 122–123
Stilul de viață adventist de ziua a șaptea (SDA), 13 Testul Shapiro–Wilk (SW) GoF, 252–253
randomizare simplă, 347 funcții source() , 123–124 foi de calcul, 87
eroarea standard a mediei (SEM), 237 rata de mortalitate standardizată (SMR), 48 teoria eșantionării statistice, 42 stratificare, 377–378
stratificarea, ajustarea de confuzie suplimentarea cu vitamina E și riscul de CHD
analiza datelor, 378
Metoda MH, 380–382
cota de cote (OR), 379
raportul de risc, 379
istoricul fumatului, 379–380
probabilitate subiectivă, 212 statistici rezumate, 232 analize de supraviețuire
teorema lui Bayes
Inferența bayesiană, 269–270
Modelul de regresie Cox, 284–288
funcția de pericol cumulativ, 269
Procedura Kaplan-Meier, 272–283
model de regresie liniară multiplă, 283–284
confuzie
studii caz-control, 376–377
studii de cohortă, 377
coliniaritate, 386–388
confuzător, 375–376
dublu orb, 377
studiu randomizat controlat, 377
vedere schematică, 376
stratificare, 377–378
dezvoltarea fazei
epidemiologic, 364–369
farmacoepidemiologie, 363–364
faza I, 363
faza II, 363
faza III, 363
faza 0/faza preclinica, 361

analiza supraviețuirii (*cont.*)
studii randomizate (RT)
analiza datelor biostatistice, 348–350
orbit, 347
cluster, 345
crossover, 345
factorial, 345–346
grup paralel, 345
randomizare, 346–347
neorbit, 347–348

diagrame de numărare, 84
semne de bifă, 166–167
Grafică cu spalier. *Vedeți* diagrama grafică a grilei cu zăbrele

studii randomizate neorb (RT), 347–348 estimare necondiționată a probabilității maxime (UMLE), 54

Segmente de cod univariate de analiză a datelor, 133 funcția reexp() , 133, 134 Mediul R , 134–135 Funcția univax() , 134

vectori în Biostatistică, 107
Diagrama Venn, 5

latime, 169

x a time series, 157
X a numeric matrix or data frame, 157

y a function of x, 157–170
Y a function of X, 157–170

z risk factor, 285

1. Urmați instrucțiunile pas cu pas date în paragrafele de deschidere ale primei sesiuni pentru a configura un mediu R. Fereastra R ar trebui să arate astfel:
>
Acum introduceți următoarele operații aritmetice; nu uitați să apăsați ENTER după fiecare intrare:
(a) 2 + 3 <Enter>
(b) 13 – 7 <Enter>
(c) 17 * 23 <Enter>
(d) 100/25 <Enter>10/25
(e) Ați obținut următoarele rezultate: 5, 6, 391, 4?

2. Iată încă câteva. (De acum înainte, promptul <Enter> va fi omis.)
(a) 2^4
(b) sqrt(3)
(c) 1i [1i este utilizat pentru unitatea complexă i , unde $i^2 = 1$.]
(d) (2 + 3i) + (4 + 5i)
(e) (2 + 3i) * (4 + 5i)

3. Iată o scurtă sesiune despre utilizarea lui R pentru a face aritmetică complexă. Doar introduceți următoarele comenzi în mediul R și raportați rezultatele:
> th <- seq(-pi, pi, len=20)
> th (a) Câte numere sunt tipărite?

```
> z <- exp(1i*th)
> z (b) Câte numere complexe sunt
tipărite?
> par(pty="s")
```

(c) De-a lungul barei de meniu din partea de sus a mediului R :

- Selectaţi şi faceţi clic stânga pe „Fereastră”.
- Deplasați-vă în jos și selectați a doua opțiune:

Dispozitiv grafic R 2 (ACTIV)

1. Efectuați următoarele exerciții elementare de aritmetică:

(a) 7 + 31
(b) 87 – 23
(c) 3,1417 × (7) ^{2}
(d) 22/7
(e) e $^{\sqrt{2}}$

În programul MPH, cinci studenți absolvenți care urmează clasa numită „Introducere

Epidemiologia” le-a măsurat greutatea (în kilograme) și înălțimea (în metri). The

rezultatul este rezumat în următoarea matrice:

În investigațiile epidemiologice, după pregătirea seturilor de date colectate pentru a fi analizate biostatistice (așa cum sa discutat în Secțiunea 3.1), primul pas este introducerea seturilor de date în mediul R. Odată ce seturile de date sunt plasate în mediul R , analiza va procesa datele pentru a obține rezultate care să conducă la creditabile

1. Folosind segmentul de cod R aici:

x 50 de 50 de numere aleatoare din distribuția normală standard.

(b) Ieșire *x* .

(c) Efectuați o analiză univariată a datelor pe *x* :

```
> x <- rnorm(1:50)
> x
> install.packages(„epibasix")
> bibliotecă (epibasix)
> univar(x)
```

> x

Un **studiu caz-control** este o clasă de studiu observațional epidemiologic. Un *studiu observațional* este unul în care subiecții caz nu sunt randomizați în grupurile expuse sau neexpuse, ci mai degrabă sunt observați pentru a determina atât expunerea, cât și statutul lor de rezultat. Starea expunerii nu este astfel determinată de cercetător.

Studiul caz-control poate fi considerat un studiu epidemiologic observațional al persoanelor cu boala de interes – grupul **de caz** – împreună cu un grup **de control** adecvat de persoane fără boală (*grupul de comparație* sau *grupul de referință*). Investigația urmărește relația potențială a unui factor de risc suspectat (sau a unui atribut al bolii) prin compararea subiecților bolnavi și nebolnavi cu privire la cât de frecvent este prezent factorul de risc în fiecare dintre cele două grupuri. Un studiu caz-control este adesea contrastat cu studiile de cohortă, în care subiecții expuși și neexpuși sunt observați până când dezvoltă un rezultat de interes.

O **cohortă** este un grup de persoane care au o caracteristică comună într-o perioadă definită (de exemplu, membrii cohortei au fost supuși unei anumite proceduri medicale , au fost expuși la un medicament sau s-au născut într-o anumită perioadă). De exemplu, grupul de oameni care s-au născut într-o anumită perioadă formează o *cohortă de naștere* : Baby Boomers (cei care s-au născut după 1946 și înainte de 1964) sunt un astfel de grup. Grupul de comparație poate fi o altă cohortă de subiecți care au avut o expunere redusă sau deloc la substanța investigată sau poate fi populația generală din care este extrasă cohorta (dar similară de altfel) și așa mai departe. În plus, subgrupurile - din *cadrul* cohortei pot fi comparate între ele.

Un **studiu de cohortă** este o formă de studiu longitudinal (în sine un tip de studiu

observațional) utilizat în investigațiile epidemiologice [1]. Este o analiză a factorilor de risc efectuată prin urmărirea unui grup de subiecți de caz care nu au boala și utilizează corelații pentru a determina riscul absolut al unui subiect de a contracta boala de interes. Este unul dintre mai multe tipuri de design de studii clinice și ar trebui comparat cu un studiu transversal. Studiile de cohortă se ocupă în general de istoriile de viață ale segmentelor de populație, precum și de persoanele individuale care constituie - aceste segmente.